Dynamics of Insertion Sequences in Bacterial Genomes

Jane Hawkey

ORCID: 0000-0001-9661-5293

Doctor of Philosophy

Department of Biochemistry and Molecular Biology & Faculty of Veterinary and Agricultural Sciences
The University of Melbourne

Submitted in Total Fulfillment of the Requirements of the Degree of Doctor of Philosophy

Abstract

Insertion sequences (IS) are small, transposable elements that encode a transposase required for their own transposition. IS are commonly found in bacteria, and are present in both chromosomes and plasmids. IS can contribute significantly to the evolution of bacterial genomes in various ways: insertion of an IS within a gene results in inactivation of that gene, IS located upstream of a gene can cause over-expression, and two IS are able to mobilise genes located between them. Currently, genomic studies of bacteria largely ignore the contribution of IS to genome evolution, as they are difficult to detect in short read sequencing data due to their repetition within genomes. To appropriately study these elements in large bacterial genomic studies, there needs to be a tool that can detect them accurately from short read data.

For this thesis, I developed a novel method for detecting IS from short read sequencing data. The resulting tool, ISMapper, was validated by looking for IS in three types of short read data: i) simulated short reads from completed genomes, ii) Illumina reads from bacteria where the finished genomes were available to compare with, and iii) Illumina reads from seven *Acinetobacter baumannii* genomes, where predicted IS sites were confirmed using PCR. In all cases, ISMapper detected all IS sites with a high degree of specificity and sensitivity.

To demonstrate the utility of ISMapper, I applied it to examine the role of IS and antibiotic resistance in multiple bacterial pathogens. ISMapper was used to determine the variable structures in the multi-drug resistance *Salmonella* Genomic Island in *Salmonella* Kentucky, finding that IS*26* was responsible for the majority of variation in this region. In *Salmonella* Typhi and *A. baumannii*, I used ISMapper to track the movement of antibiotic resistance regions and identify cryptic causes of polymyxin resistance. Each of these examples showcased the utility of using ISMapper to track the movement of IS in bacterial genomes, and how IS can contribute to the spread of antibiotic resistance.

I then applied ISMapper to explore the contribution of IS to the evolution of three *Shigella* species:, *S. sonnei*, *S. dysenteriae*, and *S. flexneri*. All three species were found to have a high burden of IS in their genomes, particularly associated with the expansion of IS*1*, however only *S. sonnei* was found to be still subject to ongoing IS expansion, whereas IS saturation appears to have been reached in *S. dysenteriae* and *S. flexneri*. Comparison of IS in the wider *E. coli* population showed that while *Shigella*-associated IS are common across *E. coli* genomes, they are constrained to a much lower copy number, even in emerging pathogenic *E. coli* lineages.

In summary, this study examined the contribution of IS to antibiotic resistance and the evolution of the pathogen *Shigella*. The framework developed in this study for analysing the dynamics of IS in *Shigella* may be used in the future for investigating the dynamics of IS in other bacterial species.

Declaration

This is to certify that:

i) This thesis, entitled "Dynamics of insertion sequences in bacterial genomes", comprises only my original work towards this Ph.D except where otherwise indicated;

ii) Due acknowledgement has been made in the text to all other material used;

iii) The thesis is fewer than 100, 000 words in length, exclusive of tables, maps, bibliographies and appendices.

Jane Hawkey B.Sc (Hons)

Department of Biochemistry and Molecular Biology

&

Faculty of Veterinary and Agricultural Sciences

The University of Melbourne

Preface

In this preface a summary is provided of the contents of each chapter in this thesis, the chapters which are included as publications are listed, and the contribution to the chapters from the co-authors and supervisors are listed.

Chapter 1: Introduction is an original overview of the background, key concepts and motivations for this thesis, with the original works cited.

Chapter 2: Introducting ISMapper is an original work that resulted in a publication in *BMC Genomics*. The author of this thesis was first author and the main contributor of work presented in this publication.

The nature and extent of my contributions to this chapter and are detailed below:

- I contributed to the design of this published study and interpretation with Kathryn E. Holt and Helen Billman-Jacobe.
- I performed the majority of software development and validation, with code contributions from David J. Edwards and Ryan R. Wick and input from Kathryn E. Holt.
- Genomic extractions and identification of ISAba1 sites in *Acinetobacter baumannii* were performed by Mohammad Hamidian and Ruth M. Hall.
- I was responsible for the planning, drafting, editing and submission of the manuscript. All co-authors edited the manuscript.

Chapter 3: Applications of ISMapper to study antimicrobial resistance is an original work. The nature and extent of my contributions are detailed below.

- Collection, genomic extraction and phenotyping of *Salmonella* Kentucky isolates was undertaken by Simon Le Hello and François Xavier-Weill.
- Collection, genomic extraction and bioinformatic analysis of *Salmonella* Typhi isolates was undertaken by Vanessa Wong *et. al.* I contributed to the insertion sequence analysis of this data by applying ISMapper.
- Collection, genomic extraction and bioinformatic analysis of the Vietnamese *Acinetobacter baumannii* isolates was performed by Mark Schultz *et. al.* I contributed to the insertion sequence analysis of this data by applying ISMapper.
- Collection, genomic extraction and phenotyping of the Singaporean *A. baumannii* isolates was performed by Tze Peng Lim *et. al.* I contributed to the bioinformatic analysis of this data.
- Collection, genomic extraction and bioinformatic analysis of the colistin resistance *A. baumanii* isolates was performed by Stephen Baker's group at the Oxford University Clinical Research Unit and Sanger Institute. I contributed to the insertion sequence analysis of this data by applying ISMapper.

Chapter 4: Dynamics of insertion sequences in *Shigella sonnei* is an original work. The nature and extent of my contributions are detailed below.

- I undertook the majority of the bioinformatics analysis with input from Kathryn E. Holt and Helen Billman-Jacobe.
- Kathryn E. Holt and Helen Billman-Jacobe contributed to the interpretation of the data and the development of the concepts presented.
- Sebastian Duchêne contributed to the development of the Bayesian modelling technique to infer the IS saturation point.

Chapter 5: Dynamics of insertion sequences in *Shigella dysenteriae* and *Shigella flexneri* is an original work. The nature and extent of my contributions are detailed below.

- I undertook the majority of the bioinformatics analysis with input from Kathryn E. Holt and Helen Billman-Jacobe.

- Kathryn E. Holt and Helen Billman-Jacobe contributed to the interpretation of the data and the development of the concepts presented.

Chapter 6: Conclusions is an original summary of the implications and significance of the work presented in this thesis.

Publications arising from candidature

Jane Hawkey, David J. Edwards, Karolina Dimovski, Lester Hiley, Helen Billman-Jacobe, Geoff Hogg, Kathryn E. Holt. Evidence of microevolution of *Salmonella* Typhimurium during a series of egg-associated outbreaks linked to a single chicken farm. *BMC Genomics* 2013 14:1 (DOI:10.1186/1471-2164-14-800)

Jane Hawkey, Mohammad Hamidian, Ryan R. Wick, David J. Edwards, Helen Billman-Jacobe, Ruth M. Hall, Kathryn E. Holt. ISMapper: Identifying insertion sequences in bacterial genomes from short read sequence data. *BMC Genomics* 2015 16:1 (DOI:10.1186/s12864-015-1860-2)

Elisabeth Njamkepo+, Nizar Fawal+, Alicia Tran-Dien+, **Jane Hawkey**+, Nancy Strockbine, Claire Jenkins, Kaisar A. Talukder, Raymond Bercion, Konstantin Kuleshov, Renáta Kolínská, Julie E. Russell, Lidia Kaftyreva, Marie Accou-Demartin, Andreas Karas, Olivier Vandenberg, Alison E. Mather, Carl J. Mason, Andrew J. Page, Thandavarayan Ramamurthy, Chantal Bizet, Andrzej Gamian, Isabelle Carle, Amy Gassama Sow, Christiane Bouchier, Astrid Louise Wester, Monique Lejay-Collin, Marie-Christine Fonkoua, Simon Le Hello, Martin J. Blaser, Cecilia Jernberg, Corinne Ruckly, Audrey Mérens, Anne-Laure Page, Martin Aslett, Peter Roggentin, Angelika Fruth, Erick Denamur, Malabi Venkatesan, Hervé Bercovier, Ladaporn Bodhidatta, Chien-Shun Chiou, Dominique Clermont, Bianca Colonna, Svetlana Egorova, Gururaja P. Pazhani, Analia V. Ezernitchi, Ghislaine Guigon, Simon R. Harris, Hidemasa Izumiya, Agnieszka Korzeniowska-Kowal, Anna Lutyńska, Malika Gouali, Francine Grimont, Céline Langendorf, Monika Marejková, Lorea A.M. Peterson, Guillermo Perez-Perez, Antoinette Ngandjio, Alexander Podkolzin, Erika Souche, Mariia Makarova, German A. Shipulin, Changyun Ye, Helena Žemličková, Mária Herpay, Patrick A.D. Grimont, Julian Parkhill, Philippe Sansonetti, Kathryn E. Holt, Sylvain Brisse, Nicholas R. Thomson, François-Xavier Weill. Global phylogeography and evolutionary history of *Shigella dysenteriae* type 1. *Nature Microbiology* 2016 1:4 (DOI:10.1038/nmicrobiol.2016.27)

+ These authors contributed equally to this work.

Other publications arising from candidature

Kathryn E Holt, Mohammad Hamidian, Johanna J. Kenyon, Matthew T. Wynn, **Jane Hawkey**, Derek Pickard, Ruth M. Hall. Genome sequence of *Acinetobacter baumannii* strain A1, an early example of antibiotic-resistant global clone 1. *Genome Announcements* 2015 3:2 (DOI:10.1128/genomeA.00032-15)

Vanessa K. Wong, Stephen Baker, Derek J. Pickard, Julian Parkhill, Andrew J. Page, Nicholas A. Feasey, Robert A. Kingsley, Nicholas R. Thomson, Jacqueline A. Keane, François-Xavier Weill, David J. Edwards, **Jane Hawkey**, Simon R. Harris, Alison E. Mather, Amy K. Cain, James Hadfield, Peter J. Hart, Nga Tran Vu Thieu, Elizabeth J. Klemm, Dafni A. Glinos, Robert F. Breiman, Conall H. Watson, Samuel Kariuki, Melita A. Gordon, Robert S. Heyderman, Chinyere Okoro, Jan Jacobs, Octavie Lunguya, W. John Edmunds, Chisomo Msefula, Jose A. Chabalgoity, Mike Kama, Kylie Jenkins, Shanta Dutta, Florian Marks, Josefina Campos, Corinne Thompson, Stephen Obaro, Calman A. MacLennan, Christiane Dolecek, Karen H. Keddy, Anthony M. Smith, Christopher M. Parry, Abhilasha Karkey, E. Kim Mulholland, James I. Campbell, Sabina Dongol, Buddha Basnyat, Muriel Dufour, Don Bandaranayake, Take Toleafoa Naseri, Shalini Pravin Singh, Mochammad Hatta, Paul Newton, Robert S. Onsare, Lupeoletalalei Isaia, David Dance, Viengmon Davong, Guy Thwaites, Lalith Wijedoru, John A. Crump, Elizabeth De Pinna, Satheesh Nair, Eric J. Nilles, Duy Pham Thanh, Paul Turner, Sona Soeng, Mary Valcanis, Joan Powling, Karolina Dimovski, Geoff Hogg, Jeremy Farrar, Kathryn E. Holt, Gordon Dougan. Phylogeographical analysis of the dominant multidrug-resistant H58 clade of *Salmonella* Typhi identifies inter-and intracontinental transmission events. *Nature Genetics* 2015 47:6 (DOI:10.1038/ng.3281)

Tze Peng Lim, Rick Twee-Hee Ong, Pei-Yun Hon, **Jane Hawkey**, Kathryn E. Holt, Tse Hsien Koh, Micky Lo-Ngah Leong, Jocelyn Qi-Min Teo, Thean Yen Tan, Mary Mah-Lee Ng, Li Yang Hsu. Multiple genetic mutations associated with polymyxin resistance in *Acinetobacter baumannii*. *Antimicrobial Agents and Chemotherapy* 2015 59:12 (DOI:10.1128/AAC.01884-15)

Mohammad Hamidian, **Jane Hawkey**, Kathryn E. Holt, Ruth M. Hall. Genome sequence of *Acinetobacter baumannii* strain D36, an antibiotic-resistant isolate from lineage 2 of global clone 1. *Genome Announcements* 2015 3:6 (DOI:10.1128/genomeA.01478-15)

Mark B Schultz, Duy Pham Thanh, Nhu Tran Do Hoan, Ryan R. Wick, Danielle J. Ingle, **Jane Hawkey**, David J. Edwards, Johanna J. Kenyon, Nguyen Phu Huong Lan, James I. Campbell, Guy Thwaites, Nguyen Thi Khanh Nhu, Ruth M. Hall, Alexandre Fournier-Level, Stephen Baker, Kathryn E. Holt. Repeated local emergence of carbapenem-resistant *Acinetobacter baumannii* in a single hospital ward. *Microbial Genomics* 2016 2:3 (DOI:10.1099/mgen.0.000050)

Sebastian Duchêne, Kathryn E. Holt, François-Xavier Weill, Simon Le Hello, **Jane Hawkey**, David J. Edwards, Mathieu Fourment, Edward C. Holmes. Genome-scale rates of evolutionary change in bacteria. *Microbial Genomics* 2016 2:11 (DOI:10.1099/mgen.0.000094)

Acknowledgements

First and foremost, I would like to extend my sincere thanks to my supervisors. To Kat, my primary supervisor, who was always patient with me as I learnt the skills required to undertake this project, and has always challenged me to do my best work. She has been a wonderful mentor, who has taken the time to ensure I have a rewarding scientific and social experience during my Ph.D. To Helen, my co-supervisor, who introduced me to the world of research during my honours year. Helen provided me with my first taste of beginner's bioinformatics, and supported me on my journey into the Ph.D program. Throughout the entirety of my Ph.D, she has provided invaluable guidance and insight. Both supervisors have been outstanding in their encouragement of me, and without them I wouldn't be where I am today.

This project would not have been possible without the support of my colleagues, both professionally and personally. I would like to thank David Edwards for his help learning the ropes in the very beginning, and allowing me to continually break his software as I learnt to program. Ryan Wick was always on hand to assist me with difficult bioinformatics problems or understanding the bash command line. Paul McAdam, Mark Schultz and Sebastian Duchene provided phylogenetics expertise whenever I needed it, for which I am deeply grateful. To everyone in the Holt lab - Zoe Dyson, Maggie Lam, Shu Mei Teo, Yu Wan, Kelly Wyres, and those mentioned above - you have all made my Ph.D experience enjoyable. Not only have you provided me with technical help when I needed it, you have been fun, kind and welcoming. I would also like to thank all those in the Inouye lab and the Centre for Systems Genomics for their support and friendship during my Ph.D.

The support of my friends and house-family has been instrumental in my success. Firstly, my housemates Cat de Burgh-Day and Franky Dillon - thank you for your friendship over the years. Throughout my Ph.D you have consistently been on hand to distract me, either by watching terrible TV shows or movies with me, boardgames, or Easter egg hunts through the

house. Danielle Ingle, Claire Gorrie and Stephen Watts, one of the best parts of my Ph.D was meeting the three of you. You were always there when I needed you, and I don't think I could have finished without you. To Alex Malone, Simone Vass, Hayley House and Dave Shaw - I appreciate all the support, friendship, and advice you have given me. I would also like to thank the friends I made during my time in the bioinformatics student group, COMBINE. Harriet Dashnow and Andrew Lonsdale have always been on hand to allow me to vent about my Ph.D or student politics. Leah Roberts was a wonderful vice-president who stepped up to the plate when my thesis prevented me from performing my COMBINE duties.

I would like to thank my extended family, but especially Pa, Nanny, Uncle John, Aunty Alison, Aunty Cathy, Uncle Conrad and Aunty Wendy, who have consistently taken a great interest in science and my education, and have always been up for a fiery debate about something scientific at every family shindig. I wish that Aunty Alison were here so she could read my thesis - I have no doubt that she would want to, even if I incurred her wrath by having the temerity to suggest that she might not find it very interesting.

Brendan, without your love and support none of this would have been possible. You provided me with the perspective I needed when things got hard, and terrible puns about my research (or anything else, for that matter) to make me laugh.

Lastly, to my parents, Ken and Linda, and my sister, Kate - thank you. You have always inspired me to think about how the world works, and continually supported me throughout my education. Your encouragement has been essential, and I wouldn't be here without you.

Table of Contents

1 Introduction **1**

1.1 The evolution of bacterial genomes 2

1.1.1 Incorporating larger DNA segments via horizontal gene transfer . . . 3

1.1.2 Genome reduction through gene loss 8

1.1.3 Swapping DNA via recombination and genome rearrangements 9

1.2 The importance of IS in bacterial genomes 10

1.2.1 Hypotheses of IS abundance in bacterial genomes 10

1.2.2 The effect of IS on bacterial evolution 12

1.2.3 IS and the development of antimicrobial resistance in bacteria 15

1.2.4 IS transposition mechanisms 17

1.2.5 Grouping of IS into IS families 20

1.3 Approaches in genomics to investigate bacterial evolution 24

1.3.1 Sequencing bacterial genomes with whole genome sequencing 24

1.3.2 Reconstructing bacterial genomes with assembly 26

1.3.3 Detecting variation in bacterial genomes using mapping 27

1.3.4 Phylogenetics: understanding evolutionary relationships between genomes 28

1.3.5 Using genomics to investigate the impact of IS on bacterial genomes . 29

1.4 *Shigella spp.* 30

1.4.1 *Shigella* pathogenesis 30

1.4.2 *Shigella* has evolved from *E. coli* 31

1.4.3 Population history and genomic studies of *Shigella spp.* 34

1.5 Aims of this project 39

2 Introducing ISMapper **41**
2.1 Introduction . . . 42
2.2 Publication . . . 44
2.3 Tools published since ISMapper's publication . . . 55
2.3.1 ITIS . . . 55
2.3.2 ISseeker . . . 55
2.3.3 ISQuest . . . 56
2.4 Conclusion . . . 57

3 Applications of ISMapper to study antimicrobial resistance **59**
3.1 Introduction . . . 60
3.2 Reconstructing AMR elements in *Salmonella* Kentucky . . . 60
3.2.1 Multi-drug resistant *S.* Kentucky ST198 . . . 60
3.2.2 Genomic study of *S.* Kentucky ST198 . . . 62
3.2.3 Methods . . . 63
3.2.4 Phylogeographic analysis of *S.* Kentucky ST198 based on whole genome SNP data . . . 66
3.2.5 Variation within the SGI in *S.* Kentucky ST198 . . . 68
3.3 The spread of antibiotic resistance in *S.* Typhi . . . 74
3.3.1 Multi-drug resistant *S.* Typhi . . . 74
3.3.2 Genomic analysis of a global collection of *S.* Typhi . . . 74
3.3.3 ISMapper analysis . . . 75
3.3.4 Conclusions . . . 75
3.4 Antibiotic resistance in *Acinetobacter baumannii* . . . 77
3.4.1 *A. baumannii* and resistance . . . 77
3.4.2 IS-mediated resistance in *A. baumannii* isolated from a Vietnamese intensive care unit . . . 78
3.4.3 IS-mediated resistance to polymyxins in *A. baumannii* . . . 80
3.4.4 Common pathways to IS-mediated polymyxin resistance . . . 86
3.5 Discussion . . . 86

4 Dynamics of insertion sequences in *Shigella sonnei* **89**
4.1 Introduction . . . 90
4.1.1 Aims . . . 90

4.2 Methods . . . 91
4.2.1 Selection of IS and creation of IS-free reference genome . . . 91
4.2.2 Detection of IS, nonsense SNPs and indels . . . 92
4.2.3 Assembly of all *S. sonnei* genomes . . . 93
4.2.4 Phylogenetic trees used in this study . . . 93
4.2.5 Ancestral reconstruction of IS insertion sites, nonsense SNPs, and indels 94
4.2.6 Modelling of changes in IS copy number over time . . . 95
4.2.7 Identification of genes under balancing or negative selection . . . 97
4.2.8 Identifying RAST categories enriched for inactivated genes . . . 99
4.2.9 Detection of colicin genes . . . 99
4.3 Results . . . 100
4.3.1 IS burden in *S. sonnei* lineages . . . 100
4.3.2 Evolutionary history of IS in *S. sonnei* . . . 104
4.3.3 Distribution of IS in the *S. sonnei* genome . . . 115
4.3.4 The role of IS in negative and balancing selection amongst *S. sonnei* genomes . . . 117
4.3.5 Diversification of gene inactivation amongst *S. sonnei* lineages . . . 122
4.4 Discussion . . . 126
4.4.1 Strengths and limitations . . . 126
4.4.2 IS in *S. sonnei* behave differently depending on lineage . . . 127
4.4.3 IS-mediated genome decay is ongoing in *S. sonnei* . . . 129
4.5 Summary . . . 131

5 Insertion sequences in *Shigella dysenteriae* and *Shigella flexneri* 133
5.1 Introduction . . . 134
5.1.1 Aims . . . 135
5.2 Methods . . . 136
5.2.1 *S. dysenteriae* data and analysis . . . 136
5.2.2 *S. flexneri* data and analysis . . . 137
5.2.3 Identification of IS in *S. boydii* genomes . . . 138
5.2.4 Identifying orthologous genes in *Shigella* . . . 138
5.2.5 IS*1* sequence analysis . . . 138
5.2.6 *E. coli* data and analysis . . . 140
5.2.7 Detection of IS in ST131, ST11 and O104:H4 *E. coli* . . . 140

5.3 Results . . . 141
5.3.1 IS distribution in *S. dysenteriae* Sd1 . . . 141
5.3.2 IS distribution in *S. flexneri* . . . 146
5.3.3 Similarities and differences between IS dynamics in each *Shigella* species 151
5.3.4 Comparing IS in *E. coli* with *Shigella* . . . 156
5.3.5 Functional impact of IS on gene inactivation in *Shigella* . . . 162
5.4 Discussion . . . 168
5.4.1 Strengths and limitations . . . 168
5.4.2 Contribution of IS to the evolution of each *Shigella* species . . . 170
5.4.3 IS*1* is expanded in each *Shigella* species, but not expanded in any *E. coli* lineage . . . 172
5.5 Summary . . . 173

6 Conclusions **175**
6.1 Development of a novel method for examining IS in bacteria . . . 176
6.1.1 ISMapper aids investigation of the maintenance and spread of AMR . 176
6.1.2 Future investigative possibilities using ISMapper . . . 177
6.2 A new framework for examining IS in *Shigella* . . . 178
6.2.1 Insights into the role of IS and *Shigella* evolution . . . 179
6.2.2 Future directions for understanding the role of IS in the evolution of *Shigella* . . . 179
6.3 Using ISMapper as a tool to examine IS dynamics in other bacterial pathogens 181

References **183**

Appendices **219**

List of Tables

3.1 Polymyxin B resistance mutations in each *A. baumannii* isolate in this study (adapted from Lim *et al.*, 2015). 84
3.2 Non-conserved IS*Aba1* insertions within each pair of colistin passaged *A. baumannii* isolates. 85

4.1 Example contigency table for odds ratio calculation. 99
4.2 IS detected in 126 *S. sonnei* genomes using ISMapper analysis. 101
4.3 Counts and proportions of IS insertion sites present in either the mrca of lineages II and III, or the mrca of lineage I, and IS insertion sites present in both mrca. 107

5.1 Genomes and accessions used for IS*1* phylogeny. 139
5.2 IS detected in 125 *S. dysenteriae* genomes using ISMapper analysis. 142
5.3 IS detected in 351 *S. flexneri* genomes using ISMapper analysis. 147
5.4 Number of IS copies in the ancestors of each lineage in *S. flexneri*, as estimated by maximum parsimony, compared to the median number of IS copies found in extant genomes of that lineage. Divergence dates are those estimated by Connor *et. al.* 147
5.5 Comparison of nucleotide and IS diversity in each *Shigella* species. For *E. coli*, the mean value across all genomes is shown, with the minimum and maximum values in brackets. 155
5.6 IS found in each pathogenic *E. coli* lineage, with their family and the mean proportion per genome. 161
5.7 Comparison of IS insertion rates and pairwise shared and non-shared pseudogenes for each *Shigella* species. 165

List of Figures

1.1 **Mechanisms of HGT and mobile elements. a,** The three mechanisms of HGT - transformation, transduction, and conjugation. Mobile elements are able to transfer from one DNA segement (eg, a plasmid) to another (eg, the chromosome). **b,** An insertion sequence, made up of a single transposase gene (*tnp*) and IR. **c,** The two different types of transposon, composite and non-composite. Composite transposons are genes (light orange box) flanked by two IS. Non-composite transposons have more complex transposase structures, but genes (light orange box) can still be carried within them. This non-composite transposon structure is typical of the Tn*3*-subgroup transposons. **d,** Integrons, made up of *int* and *attI* genes, capture gene cassettes, which integrate next to the *attI* sequence. The captured genes are expressed via a promoter on the integron, between *int* and *attI*. **e,** An example of a genomic island. Genomic islands are often found downstream of tRNA genes, and usually create DR upon insertion. They frequently have integrase genes, and can carry IS or other small mobile elements. 4

1.2 **Different IS transposition mechanisms.** Figure adapted from Siguier *et al.*, 2015.[72]. **a,** copy-out paste-in mechanism. Blue boxes show the IS and the grey lines are the donor DNA molecule. Maroon arrow shows where the transposase cuts. In this transposition mechanism, the bottom strand is cut and then circularises itself, with the red bar indicating where the cut site is. DNA replication then occurs to repair the single strand break, repairing the IS and creating a double stranded circular DNA molecule which can then insert itself into a new DNA molecule (purple lines). **b,** one type of cut-and-paste. Blue boxes show the IS, with maroon arrows indicating the transposase cut sites, some of which occur 2 bp within the insertion sequence (green boxes). The IS is then excised from the donor DNA molecule (grey lines) and can then 'paste' itself into a new DNA molecule (purple lines). **c,** a second cut-and-paste mechanism, with a single cut on each strand (maroon arrows). The IS then forms hairpin structures at either end of the IS molecule, before inserting itself into a new DNA molecule (purple lines). 19

1.3 **Steps in the evolution of *Shigella* from its ancestor, *E. coli*.** 33

1.4 **Bayesian maximum clade credibility phylogeny for *S. sonnei*.** Figure adapted from Holt *et al.*, 2012[250]. Branches defining major lineages are shown in bold (each with 100% posterior support). Pie charts indicate maximum-likelihood estimates for geographic origin of major nodes. Divergence dates (median estimates and 95% HPD) shown for major nodes. . 35

1.5 **Population structure of *S. dysenteriae*.** Figure adapted from Njamkepo *et al.*, 2016.[260] **a,** Maximum likelihood phylogeny of 235 *S. dysenteriae* genomes. Tips of the tree are coloured by continent, as per legend. Segments of tree are highlighted by lineage. **b,** Bayseian phylogeny of 125 *S. dysenteriae* genomes. Branches are coloured by lineage, with dates of emergence at each major node. Purple squares show intercontinental transmission events. 37

1.6 **Maximum likelihood phylogeny for *S. flexneri* isolates including serotypes 1–5, X and Y produced from the results of mapping sequence reads against the genome of *S. flexneri* 2a strain 301, with recombination removed.** Reproduced from Connor *et al.*, 2015.[264] Original legend: "Phylogenetic groups (PGs) determined by Bayesian analysis of population structure clustering are boxed within dotted lines, with the geographic and serotype composition of isolates in each PG being inlaid as pie charts." 38

3.1 **Mutation rate estimates for real and randomised tip dates in *S.* Kentucky**. First column, real mutation rate. Subsequent columns show mutation rate when tip dates are randomised. Black circles are the mean rate estimated by BEAST, with error bars showing 95% highest posterior density (HPD). 65

3.2 **Phylogeographic reconstruction of *S* Kentucky ST198**. Phylogeny is maximum clade credibility tree estimated from BEAST. Nodes and branches are coloured by region as per map. SGI type in each isolate indicated by symbol as per legend. Arrows indicate branches where different SGI types were acquired. Mutations in codons 83 and 87 in *gyrA* shown in purple, and mutations in codon 80 of *parC* shown in pink. 67

3.3 **Presence of SGI backbone genes and IS*26* copy number in each *S.* Kentucky genome.** Left, maximum clade credibility tree estimated from BEAST, with branches and tips coloured by region as per map. Country of isolation is listed next to each genome. Presence of SGI genes are shown in heatmap; red, resistance region; light blue, flanking chromosomal genes (*trmE* and *yidY*); dark blue, SGI backbone genes; grey, partial gene deletion. Orange bar chart shows IS*26* copy number for each genome, estimated by ISMapper. 70

3.4 **Examples of the variation found with each SGI1 type found in *S.* Kentucky.** Genes are represented by arrows and coloured by type as per legend. Contig breaks are shown by black bars, ISMapper hits by purple bars. 71

3.5 **Presence of antibiotic resistance genes within each genome of *S.* Kentucky.** **a,** Phylogenetic tree of ST198 MDR clone with tips coloured by region. **b,** Resistance phenotype with drugs coloured by class as per legend. **c,** Symbols represent SGI type in isolate, as per legend. Resistance genes present on the SGI, coloured by the class of antibiotic they confer resistance to. **d,** Plasmid replicon detected in each isolate, with resistance genes listed that are present on that plasmid. Gene names in brackets indicate unknown location. Gene names are coloured by antibiotic class they confer resistance to. 72

3.6 **IS*26* copy number in each genome.** *x*-axis, year genome isolated; *y*-axis, IS*26* copy number. Points are coloured by SGI type, as per legend. 73

3.7 **Phylogeny of H58 *S.* Typhi isolates**. Branches coloured by region. Innermost ring, red, shows number of resistance genes found on the Tn*2670*. Second ring shows plasmid replicon type for each isolate. Third ring shows presence of IS*1* in *cyaA* and outer ring shows Tn*2670* insertion site. Figure adapted from Wong et. al., 2015.[276] 76

3.8 **Location of transposon insertion sites in *cyaA* and *yidA***. Top, CT18 reference at *cyaA*. Next two panels show transposon inserted in *cyaA* and then *yidA*. Bottom, CT18 reference at *yidA* site. Figure reproduced from Wong et. al., 2015.[276] 77

3.9 **Core genome phylogeny for GC2 *A. baumannii* in the ICU.** BEAST maximum clade credibility tree; shading indicates the period during which imipenem was used for the empirical treatment of VAP in the ICU. Isolate labels are coloured to indicate source: red, VAP; blue, asymptomatic carriage. Node bars indicate 95 % HPDs for divergence dates; node labels and branch line thickness indicate posterior support. The two main lineages (1 and 2) and five imipenem-resistant subclades (A–E) are labelled, arrows indicate inferred *oxa23* carbapenemase acquisition events: red, Tn*2006*; orange, Tn2008VAR. *Oxa23* gene copy number and MICs for imipenem are indicated on the right. Figure reproduced from Shultz et. al., 2016.[69] 79

4.1 **Burden of IS in *S. sonnei*. a,** Box plots showing overall proportion of each IS within *S. sonnei* genomes. IS family is indicated across top. Orange dots indicate the proportion of that IS within the most recent common ancestor (mrca) of *S. sonnei*. **b,** Number of strain-specific insertions per IS, with bars coloured by IS family, and IS family printed across top. **c,** PCA plot of IS profiles (as shown in Figure 4.2) for each genome. Points are coloured by lineage as per legend. **d,** Box plots of total IS copy number within each lineage. **e,** Phylogeny of *S. sonnei* with lineages labelled and coloured. Marked nodes indicate year of most recent mrca with number of IS copies indicated underneath in brackets. **f,** Stacked bar plots of IS copy number in each genome, and mrcas, coloured by IS as per legend. 102

4.2 **Presence of IS insertion sites in individual *S. sonnei* genomes.** Each row represents a genome, corresponding to the tip in the BEAST tree (left). Each column respresents a specific IS insertion site; coloured blocks indiciate presence of the IS insertion in each genome. Columns are clustered by IS, not ordered by position within the genome. 103

4.3 **Dynamics of IS with *S. sonnei*. a,** Scatterplot of IS copy number in all extant genomes, with points coloured by lineage. Thick horizontal lines on y axis indicate total IS copy number in each mrca, as per legend. R^2, p and slope values for each regression model indicated in coloured text above the x axis. **b,** Phenogram of tree showing total IS copy number inferred at each node on the tree, except for the mrca of *S. soneni*. Branches coloured by lineage as per legend. Dashed lines indicate the two possible reconstructions from the lower and upper bounded root mrca copy numbers, with lines coloured as per legend. 108

4.4 **Relationship between gain and loss events for each IS inferred across the *S. sonnei* phylogeny, and the relationship between gain events and strain-specific insertions, for each IS. a,** Scatterplot showing the relationship between gain and loss events, for each IS, inferred across the phylogeny, using maximum parsimony ancestral state reconstruction (see Methods 4.2.5). Estimates of gain and loss events do not include events at the mrca of *S. sonnei*, the mrca of lineage I, or the mrca of lineages II and III. Dots are coloured by IS as per legend. Grey line shows x=y. **b,** Scatterplot showing total number of gain events against the number strain-specific insertions, for each IS. Dots are coloured by IS as per legend. Black line shows the linear relationship between the number of gain events and strain-specific insertions. 109

4.5 **Mutation rates in each lineage of *S. sonnei*.** Medians (black dots) and 95% HPD intervals (black lines) for mutation rates in each lineage as estimated by BEAST (lineage I, blue; lineage II, purple; lineage III, green). First column in each plot shows the mutation rate estimated using the real isolation dates, with the remaining ten columns showing mutation rates estimated with randomised dates. 110

4.6 **Rates of gain in each *S. sonnei* lineage.** Mean number number of IS gain events per branch, with branches binned by decade, across 100 random trees. Points and lines are coloured by lineage as per legend. Dashed lines indicate upper and lower interquartile range. Light shaded region (prior to 1900) indicates less confident estimates of gain rate. 111

4.7 **Modeling of IS saturation point in *S. sonnei* genomes.** Points show the IS copy number of each node and tip in the *S. sonnei* phylogeny against relative time, with points coloured by lineage as per legend. Lineage II and III points are offset in time relative to lineage I. Black line is the logistic curve using parameters estimated by the modeling, with grey dotted lines showing the 95% credible interval of the curve. Insert shows enlarged section of scatterplot, focusing on the lineage points. 112

4.8 **Distributions of prior and posterior values for each parameter in the model.** Top row – prior distributions for L, k, x0 and er. Second row – posterior distributions for L, k, x0 and er. Third row – prior distributions for x_1, x_2 and x_3. Final row – posterior distributions for x_1, x_2 and x_3. 113

4.9 **Trace plots for all parameters in the model.** 114

4.10 **Density of IS insertion sites around the *S. sonnei* 53G chromosome.** Circular map of the 53G chromosome. From outer-most to inner-most ring: IS density in lineage III; IS density in lineage II; IS density in lineage I; density of IS*1* in all lineages; density of all IS except IS*1* in all lineages; MLST genes. Grey shading indicate IS hotspot regions, annotated with the coding sequences within these regions. 116

4.11 **Intergenic IS insertions in *S. sonnei*. a,** Histogram of number of insertions, per base, per gene. Intergenic rate of insertion is shown by the red line, 2x intergenic rate of insertion is shown by the blue line. **b,** Insertions per base, per gene for each gene in the *S. sonnei* genome, arranged by genome position on the x axis. Grey dots indicate genes interrupted by IS at less than the 2x intergenic rate (dashed blue line) shown in panel **a**. All other genes above the 2x intergenic rate are coloured by the number of IS insertions found within them. 120

4.12 **Comparison of *btuB* interruptions and presence of colicin in each *S. sonnei* genome.** Light blue bars indicate an uninterrupted copy of *btuB*, dark blue bars show presence of an interruption in *btuB* or its promoter by either IS or mutation. Locations of *btuB* interruptions are shown in cartoon above phylogeny, with the promoter marked as a black arrow. Red and orange bars indicate presence of the colicin E1 toxin or immunity gene. Purple bars indicate the presence of the colicin E3 toxin or immunity gene. 121

4.13 **Accumulation of inactivated genes in each *S. sonnei* lineage. a,** Bar plot with total height of bar illustrating the median number of inactivated genes in each lineage. Blue, number of genes inactivated by IS; purple, number of genes inactivated by both IS and mutation; red, number of genes inactivated by mutation. **b,** Phenogram showing total number of genes inactivated (by either IS or mutation) at each node, with branches coloured by lineage. Dark orange circle and dashed lines shows the maximum number of genes that could have been interrupted in the mrca. Light orange circle and dashed lines shows the minimum number of genes that could have been interrupted in the mrca. **c,** Box plots showing differences in pseudogenes between pairs of strains from within the same lineage and between lineages, broken down by mutation type. **d,** Number of inactivated genes in each genome. Left, phylogenetic tree of *S. sonnei*, with branches coloured by lineage. Right, bar plot showing total number of inactivated genes in each genome, coloured by inactivation type as per legend. 125

4.14 **Two different hypotheses for the burden of IS in the mrca of *S. sonnei*. a,** Lineages II and III rapidly accumulate IS, while lineage I accumulates IS at a steadier rate. **b,** Lineages II and III accumulate IS at a steady rate, while lineage I rapidly loses IS, then gains additional IS copies later at a more steady rate. . . 128

5.1 **Burden of IS within *S. dysenteriae*. a,** Box plots of IS proportion for each IS within the genome, with IS family indicated across top. **b,** Bar plots showing the number of strain-specific insertion sites for each IS, with IS family indicated across top. **c,** PCA of IS profiles for each genome, illustrating that lineages can be separated based on their IS profile. **d,** Box plots showing range of copy numbers within each lineage (lineage I not shown as it consists of a single isolate). **e,** Maximum clade credibility tree from Njamkepo *et. al.*[260], with branches coloured by lineage. Dates indicate time of ancestor at major nodes, numbers in brackets show inferred IS copy number at that node. **f,** Bar plots showing total copy number of each IS for each isolate in the tree, with segments coloured by IS as per legend. 143

5.2 **All IS insertion sites found in *S. dysenteriae* against the maximum likelihood phylogeny.** Heatmaps of IS position are clustered to highly patterns between lineages. Heatmaps are divided into each IS type, and coloured shades of the same colour to indicate membership to the same IS family. 144

5.3 **Scatterplot of total IS copy number in each genome, against the year genome was isolated**. Dots coloured by lineage, as per legend. Lines show linear regressions for each lineage. 145

5.4 **Burden of IS within *S. flexneri*. a,** Box plots showing proportion of each IS within the genome, with IS family indicated across top. **b,** Bar plots showing the number of strain-specific insertion sites for each IS, with with IS family indicated across top. **c,** PCA of IS profiles for each genome, showing that lineages can be separated based on their IS profile. **d,** Box plots showing range of IS copy numbers within each lineage. **e,** Maximum likelihood phylogeny of all 351 *S. flexneri* genomes from Connor *et. al.*[264], with branches coloured by lineage. **f,** Bar plots showing total copy number of each IS for each genome, with bar segments coloured by IS as per legend in **e**. 148

5.5 **Scatterplot showing total IS copy number in each genome, against the year genome was isolated.** Dots are coloured by lineage as per legend, with lines showing a best fit linear regression for each lineage. 149

5.6 **All IS insertion sites found in *S. flexneri* against the maximum likelihood phylogeny.** Heatmaps of IS position are clustered to highly patterns between lineages. Heatmaps are divided into each IS type, and coloured shades of the same colour to indicate membership to the same IS family. 150

5.7 **Comparison of burden for five IS found in each *Shigella* species.** Proportion of five common IS, IS*1*, IS*2*, IS*4*, IS*600* and IS*911* in all three *Shigella* species. Median IS copy number in each species is labelled in black at the top of each bar. 154

5.8 **Burden and proportions of the five common IS in *Shigella* and *E. coli*** Line graph indicating estimated burden of each of the five common IS in all three *Shigella* species and several *E. coli* STs. Burden is estimated as the ratio of mean read depth of IS against mean read depth across MLST genes. 158

5.9 **Comparison of IS content in the three *Shigella* species and three pathogenic *E. coli* lineages.** **a,** Box plots showing IS copy number in each *Shigella* species and each pathogenic lineage of *E. coli*, estimated using ISMapper. **b,** IS copy numbers for each pathogenic *E. coli* lineage, estimated using ISMapper. Boxplots illustrating IS copy number for all IS detected in ST131, ST11 and O104:H4. Highlighting indicates IS found in at least one *Shigella* species, coloured by IS family, as per legend. 159

5.10 **Maximum likelihood phylogeny of 827 IS*1* sequences sourced from ten *E. coli* reference genomes and five *Shigella* genomes.** The phylogeny is midpoint rooted. Branches have been collapsed to highlight relationships between groups. Branches are coloured by the species the IS*1* sequence was extracted from. Pink - *E. coli*; green - *S. sonnei*; purple - *S. dysenteriae*; orange - *S. flexneri*. 160

5.11 **Number of genes inactivated in each *Shigella* species.** Colours inside bar plots indicate the cause of gene inactivation, either IS interruption, mutation, or both. 166

5.12 **Comparison of number of genes inactivated in each *Shigella* species.** 167

Abbreviations

The following abbreviations have been used throughout the thesis:

AMR antimicrobial resistance

bp base pair

CI confidence interval

DR direct repeats

EHEC enterohemorrhagic *Escherichia coli*

ESKAPE *Enterococcus faecium*, *Staphylococcus aureus*, *Klebsiella pneumoniae*, *Acinetobacter baumannii*, *Pseudomonas aeruginosa* and *Enterobacter spp.*

GC1 global clone 1

GC2 global clone 2

HGT horizontal gene transfer

HIV human immunodeficiency virus

HPD highest posterior density

IR inverted repeats

IS insertion sequence

kbp kilobases

LPS lipopolysaccharide

Mbp megabases

MCC maximum clade credibility

MCMC Markov chain Monte Carlo

MDR multi-drug resistant

MLST multi-locus sequence type

mrca most recent common ancestor

OR odds ratio

orf open reading frame

PCA principal component analysis

sd standard deviation

SGI *Salmonella* Genomic Island

SNP single nucleotide polymorphism

ST sequence type

T3SS type III secretion system

TSD target site duplication

UPEC uropathogenic *Escherichia coli*

WGS whole genome sequencing

Chapter 1

Introduction

1.1 The evolution of bacterial genomes

Evolution is a vital process by which all life changes and adapts. Bacterial genome evolution is complex and occurs via many mechanisms, on both small and large scales. All of these mutational processes are subject to selection pressures from the environment. Selection can either be positive, where the change is beneficial to the organism and so is maintained, negative, where the change is detrimental to the survival of the organism, or neutral. Interaction of mutational processes and selection combine to facilitate adaptation of bacteria to new environments.

Small changes to the genome occur when errors are made during DNA replication or repair of DNA damage. Point mutations caused by base substitutions or insertions or deletions (indels), affect only a few base pairs. Single base changes are commonly referred to as single nucleotide polymorphisms (SNPs). SNPs occurring in both coding and non-coding regions. There are two possible outcomes from the creation of a SNP in a coding region: a synonymous, or silent change, where the codon changes but the amino acid does not change; or a non-synonymous change, where both the codon and amino acid change. A non-synonymous mutation can also inactivate a gene if it creates a premature stop codon. If an indel occurs in a coding region, it can alter the reading frame of the gene and generate a frameshift mutation. SNPs or indels in non-coding regions can still alter gene expression by changing binding sites or altering regulation signals. Some bacteria, such as the pathogen *Mycobacterium tuberculosis*, evolve primarily using these types of point mutations[1].

Large changes in the genome can occur through the introduction of new DNA entering the genome via horizontal gene transfer (HGT). Several different types of mobile elements, such as insertion sequences, transposons, integrons, phage, genomic islands, and plasmids can enter the genome through HGT. Other large changes include recombination, where homologous segments of DNA are exchanged, or genome rearrangements. Genetic material can be inherited either vertically (by descent to daughter cells) or horizontally (genetic exchange between cells).

1.1.1 Incorporating larger DNA segments via horizontal gene transfer

Horizontal gene transfer can take place via three mechanisms: conjugation, transduction or transformation (Figure 1.1a). Conjugation occurs when a donor bacteria cell connects to a recipient cell via a sex pilus. DNA is transferred to the recipient through the passage formed by the pilus[2] (Figure 1.1a). Transduction occurs when a bacteriophage (a virus that infects bacterial cells, also called a phage) packages some of the host bacterial DNA inside its capsid (see section 1.1.1.4)[2]. The phage can transfer the bacterial DNA if it infects a new cell (Figure 1.1a). There are two types of transduction – generalised, where any part of the bacterial chromosome can be transferred; or specialised, where only genes from a certain chromosomal location can be transferred[2]. The third HGT mechanism is transformation, the uptake of DNA into a cell from the environment[2] (Figure 1.1a). All three of these HGT methods can transfer novel DNA between bacterial cells. Most HGT is selectively neutral, however, sometimes a transfer event can provide increased fitness to the host and cause a selective sweep for the new DNA[3].

Some mobile elements are DNA regions that encode proteins that promote their ability to move into new genetic contexts through HGT. There are four main types of mobile element: class I, insertion sequences (IS), as well transposons that consist of IS elements at either end, and only need one protein for transposition; class II, complex transposons with inverted repeats that require multiple proteins for transposition; class III, phage elements which are transposable; and class IV, which contains all other mobile elements including integrons[4]. In addition to these smaller mobile elements, larger mobile elements also exist, including genomic islands and plasmids.

1.1.1.1 Insertion sequences

IS are compact mobile elements, ranging from 800 bp to 2,500 bp in size. They typically contain one to two open reading frames that encode proteins for their own transposition, called a transposase[5] (Figure 1.1b). The transposase recognises the inverted repeats (IR) at the ends of the IS and then extracts and moves the IS to a new location[6]. The ends of the IS usually feature inverted terminal repeats that react with the transposase during IS excision. These elements are discussed in greater detail in section 1.2.

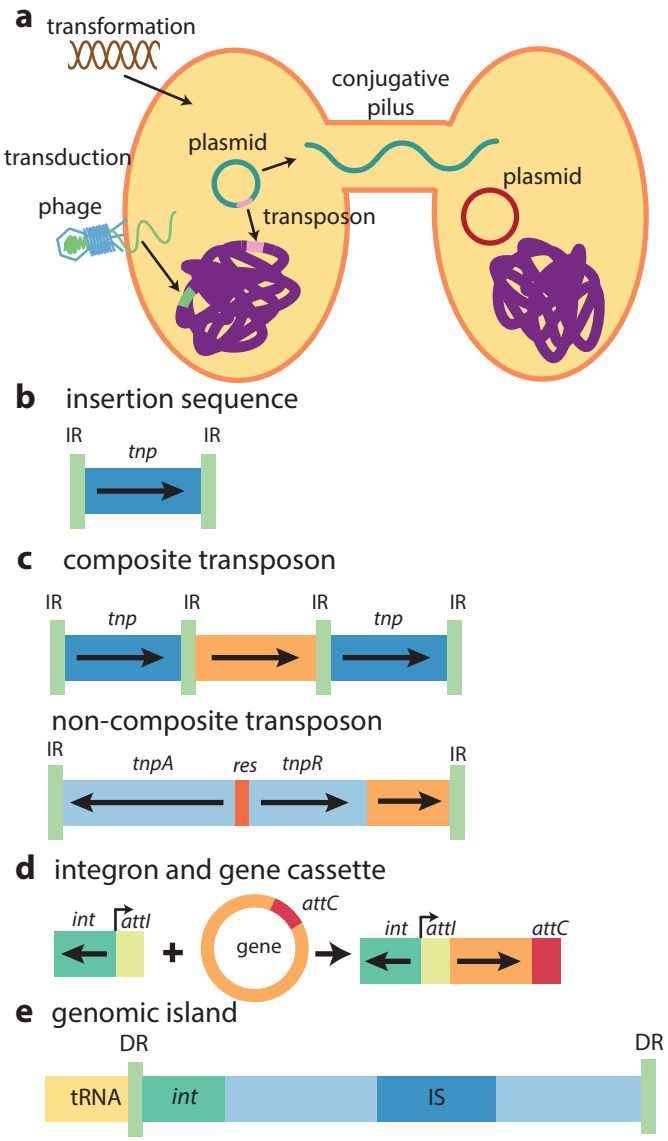

Figure 1.1: Mechanisms of HGT and mobile elements. **a,** The three mechanisms of HGT - transformation, transduction, and conjugation. Mobile elements are able to transfer from one DNA segement (eg, a plasmid) to another (eg, the chromosome). **b,** An insertion sequence, made up of a single transposase gene (*tnp*) and IR. **c,** The two different types of transposon, composite and non-composite. Composite transposons are genes (light orange box) flanked by two IS. Non-composite transposons have more complex transposase structures, but genes (light orange box) can still be carried within them. This non-composite transposon structure is typical of the Tn*3*-subgroup transposons. **d,** Integrons, made up of *int* and *attI* genes, capture gene cassettes, which integrate next to the *attI* sequence. The captured genes are expressed via a promoter on the integron, between *int* and *attI*. **e,** An example of a genomic island. Genomic islands are often found downstream of tRNA genes, and usually create DR upon insertion. They frequently have integrase genes, and can carry IS or other small mobile elements.

1.1.1.2 Transposons

Transposons are linear pieces of DNA that include a transposase and IR at each end. Transposons can be divided into three main groups: composite transposons, in which two transposases (usually IS)[7] coordinate to move the genetic material between them; non-composite transposons that are unrelated to composite transposons and are exemplified by Tn*3* or Tn*7*; and Mu type transposons, which are phage (discussed in section 1.1.1.4)[8] (Figure 1.1c).

A typical transposase identifies the terminal IR at each end of the transposon, which is between 10 and 40 bp long and divided into two functional domains[9]. In the terminal IR, one domain contains only a few base pairs and is involved in the cleavage reaction, whilst the second domain is further inside the IR and is responsible for binding the transposase[9]. However, some transposons have more complicated transposition mechanisms. For example, Tn*7* (a non-composite transposon) encodes five different transposases that together regulate two overlapping recombination pathways[10]. These transposases also determine the site of integration for the transposon and how frequently transposition occurs[10].

During transposition, some transposases will replicate the element as part of the transposition process, whilst others are excised from the genome and relocate elsewhere[8]. When moving to a new location, there are a variety of possible integration sites, depending on the transposase. Some transposases, such as in the Tn*10* transposase, show preference for a 7 bp symmetric sequence; however, others show a preference for sequences which resemble their IR[9]. Nonetheless, many transposases seem to exhibit a low target specificity[9].

1.1.1.3 Integrons

Integrons are not themselves mobile, but have the ability to become mobile with the assistance of transposons[11]. They can be carried on plasmids or within the chromosome[12]. Class I integrons are the most common and have been found in multiple species[13], and are frequently associated with multi-drug resistance (MDR) regions, especially in pathogenic bacteria[14]. However, class I integrons, which contain no resistance genes, have also been found in the Betaproteobacteria and contain no resistance genes at all, suggesting that Betaproteobacteria may contain the prototype of these elements[15].

Integrons are able to capture gene cassettes via site-specific recombination[16,17] (Figure 1.1d). The integron structure consists of an *intI* gene, of which there are over 100 types, which is required for integration, and a recombination site, *attI*, which usually lies upstream of *intI*[18]. The *intI* gene encodes a recombinase that belongs to the tyrosine recombination family[18]. The *int/att* unit which makes up an integron is different to the *int/att* units found in phage or genomic islands, as they are not themselves mobile. Rather, these units facilitate the insertion of mobile gene cassettes[18]. To capture a gene cassette, the gene cassette must have a repeat sequence, *attC*, which recognises the *attI* site on the target integron and integrates next to it[13,19] (Figure 1.1d). Gene cassettes usually contain a single open reading frame and no promoter, and expression of genes in the cassette is driven by a promoter in the integron[19,20] (Figure 1.1d).

1.1.1.4 Bacteriophage

Bacteriophage are viruses that infect bacterial cells. Their genomes can be either single or double stranded DNA or RNA, and contain genes required for their structural proteins, genome packaging proteins, and proteins for hijacking host cell processes. Lytic phage quickly replicate and lyse the host cell, however temperate phage are able to integrate into the host chromosome via a process known as lysogeny[21]. During lysogeny, the phage genome replicates along with the host chromosome and becomes an endogenous phage, or prophage, which lies dormant but may later activate and lyse the cell[22]. Some phage genomes do not integrate and exist within the bacterial cell as small, autonomous linear or circular plasmids[20]. These latent phage can then be later reactivated by some stimulus. During phage replication, phage can accidentally package some of their host's DNA into their capsid. Although they are limited in the amount of DNA that can be physically packaged into the capsid, phage can still transport sections of host DNA into new bacterial cells through transduction[22].

1.1.1.5 Genomic islands

Genomic islands often have a large impact on bacterial evolution due to the wide variety of functions they can provide, including virulence, antibiotic resistance, and catabolic pathways. Genomic islands are large segments of DNA that range in size from 10 kbp to 200 kbp[23] (Figure 1.1e). Genomic islands have a different nucleotide composition to the the rest of the

chromosome, enabling their detection within genomes[24]. They usually insert at the 3' end of tRNA genes[25]. Site specific recombinases, similar to phage, mediate insertion into the genome[26,27]. As a result of integration, genomic islands are often flanked by perfect direct repeats (DR) that are used as recognition sites for later excision[28]. Excision is mediated by an excisionase, and once excised, the island closes to form a circular piece of DNA ready for transfer to another cell[25,29]. Sometimes genomic islands are able to self mobilise, but there are examples of genomic islands requiring the assistance of other mobile elements, including plasmids or phage, to move to new locations[30]. During mobilisation, they can transfer additional host chromosomal DNA to their new location[31]. Some genomic islands contain similar genes to plasmid encoded transfer systems, suggesting that these plasmid transfer genes may be the origin of genomic island movement[32]. One possible mechanism for genomic island evolution theorizes that if a plasmid integrates into a chromosome, subsequently loses its origin of replication (ori) and replication genes, it can become a genomic island[33]. Genomic islands are hot spots in the chromosome for integration of additional mobile elements, without disrupting expression of the host genome[33] (Figure 1.1e).

Due to the large number of genes being inserted into the chromosome in a single event, the acquisition of a genomic island generates adaptation that is dramatic and fast[34]. For example, the high pathogenicity island in *Yersinia* encodes the siderophore system yersiniabactin, allowing the bacteria to scavenge iron from its environment, causing a more virulent phenotype[35]. This pathogenicity island has since been found in several other members of the Enterobacteriaciae, including *Escherichia coli*, *Klebsiella* and *Citrobacter*[36]. In *Wolinella succinogenes*, a genomic island encodes the proteins required for nitrogen fixation, expanding *W. succinogenes*' possible ecological niches[37]. Several bacteria contain genomic islands that include antibiotic resistance genes, such as the *Salmonella* Genomic Island (SGI) in *Salmonella enterica* and *Proteus mirabilis*, the *Shigella* Resistance Locus (SRL) in *Shigella*, and SXT in *Vibrio cholerae*[38–41]. In *Bacillus cereus*, the acquisition of a single genomic island not only gives this species a siderophore system, but also antibiotic resistance and the ability to generate bacteriocins[42].

1.1.1.6 Plasmids

Plasmids are mobile segments of double stranded DNA that are usually covalently closed and circular, but linear plasmids also exist, such as pBSSB1 in *S. enterica*[20,43]. Plasmids contain a

set of backbone genes that encode proteins required for their replication. Partitioning genes, that ensure the plasmid is correctly partitioned between daughter cells, determine the plasmid incompatibility (Inc) type, as two plasmids with the same machinery are unable to inhabit the same host cell[44,45]. Plasmids commonly carry genes to create conjugative pili, allowing them to transfer from one cell to another[46]. During conjugative plasmid transfer, some plasmids have been shown to occasionally transfer host chromosomal DNA[47]. Some smaller plasmids require the presence of a larger conjugative plasmid in order to transfer to a new cell[48]. In addition, plasmids frequently carry accessory genes that are not essential for cellular function, but often provide a selective advantage to the bacterial host, allowing the plasmid to remain inside the cell[20].

Plasmids do not form a part of the chromosome, but as discussed in section 1.1.1.5, they can occasionally integrate into the chromosome and lose their replicative abilities. R factors are a particular type of complex plasmid that are good at transmitting between various Gram-negative bacteria and regularly contain antibiotic resistance genes[7,49]. The antibiotic resistance genes are often contained in specific regions that are comprised of mobile elements, such as transposons and integrons, just like genomic islands. These regions frequently carry genes conferring resistance to more than one type of antibiotic, forming multi-drug resistance (MDR) plasmids. Plasmids can exist in evolutionary distinct pathogens that have direct or indirect contact with each other, either in the environment or within hosts, causing transmission of MDR regions[50]. *S. enterica* has been shown to contain many examples of MDR plasmids, such as the IncHI1 plasmids in *Salmonella* Typhi, which have contributed to the global spread of antibiotic resistant *S.* Typhi[51,52].

1.1.2 Genome reduction through gene loss

Gene loss and pseudogene formation also play a role in bacterial adaptation. Loss of genes can be achieved through small changes that subtly alter proteins, changing how a gene is regulated, or preventing expression. Gene loss can also occur through large scale processes, such as gene deletion, or interruption of genes by incoming DNA.

Protein-coding genes can be inactivated through a variety of mechanisms, including premature stop codons caused by SNPs, indels in genes causing frameshift mutations, insertion of mobile elements, or recombination of identical IS elements[53]. Gene inactivation is

commonly observed when bacteria adapt to a new ecological niche. For example, host adapted pathogens regularly inactivate metabolic pathways, as these compounds can often be scavenged from the host itself[54]. Bacteria have a deletional bias in their genomes, so genes which have been inactivated are continually eroded until they are deleted, resulting in complete gene loss[55,56].

Gene loss and inactivation frequently occur when a bacterial species undergoes a population bottleneck. Unlike in larger populations, where mildly deleterious mutations are often eliminated, in small populations, where genetic variation is smaller, mutations can become easily fixed[57]. During population bottlenecks, genes that were required for the survival of the bacteria in its new niche are inactivated; however useful, but non-essential, genes can also be inactivated[58]. *Yersinia pestis* is a bacterial pathogen where gene loss has played an important role in its evolution from *Yersinia pseudotuberculosis*. Several genes required for pathogenicity and host interaction in *Y. psuedotuberculosis* were lost in *Y. pestis*[59]. These genes were were required for spread via the fecal-oral route, but as *Y. pestis* spreads via a flea vector, they were no longer required in *Y. pestis*[59]. In addition, all of the motility genes in *Y. pestis* have been inactivated, likely to prevent recognition by the host immune system[60].

The process of pseudogene formation during host adaptation has also been observed in *Salmonella* Gallinarum and *Salmonella* Pullorum, which cause a typhoid-like illness in galliform birds. In these pathogens, many genes involved in anaerobic metabolism have been inactivated as part of the transition from a generic intestinal pathogen to an invasive, host-adapted pathogen[61,62]. Additionally, pseudogene formation has been a crucial step in the convergent evolution of *S.* Typhi and *Salmonella* Paratyphi A, where loss of functional genes has enabled both species to adapt to humans and cause typhoidal fever[63].

1.1.3 Swapping DNA via recombination and genome rearrangements

Recombination is defined as the acquisition of DNA from a donor cell into a recipient cell, and occurs independently of cell division (unlike recombination in eukaryotes)[64]. There are two types of recombination - homologous recombination, where homologous sections of DNA are switched out or replaced; and non-homologous recombination, where new DNA is integrated into the genome (more commonly known as HGT, see section 1.1.1)[64]. Rearrangement of DNA within a cell can also occur, and this is frequently mediated by mobile elements[65]. The expansion

of transposable elements such as IS are often followed by large genome rearrangements, as has been the case for multiple species of *Yersinia*.

During homologous recombination, divergent DNA can be integrated into the genome through targeting conserved genes, or identical copies of the same IS, flanking the divergent region[66]. There are several examples of this type of homologous recombination. In *Staphylococcus aureus*, large replacements of ~10% of the chromosome founded two new lineages[67]. Capsular exchange, mediated by recombination, is a common phenomenon in *Acinetobacter baumannii* and *Klebsiella pneumoniae*[68,69]. In *Streptococcus pneumoniae*, the capsule is a vaccine target, and recombination of this region has lead to capsule switching and decreased vaccine efficacy[70]. Within *E. coli*, transfer of sex factor F into the chromosome from the plasmid is mediated by identical copies of IS*2* or IS*3*[66]. Recombination and genome rearrangement are both large scale genome changes and facilitate fast adaptation[71].

1.2 The importance of IS in bacterial genomes

The primary focus of this thesis is how IS contribute to the evolution of bacterial genomes. IS are found in both chromosomes and plasmids, and contribute to the evolution of bacterial genomes in a myriad of ways. They are frequently involved in the mobilisation of genes, either by forming compound transposons or carrying genes as passenger genes[72]. This has important implications for the evolution of bacterial pathogens. Examples of genes IS can mobilise include antibiotic resistance genes and virulence genes. Additionally, IS can influence the expression of genes by either interrupting them or upregulating them, further adding to the range of phenotypes IS can alter.

1.2.1 Hypotheses of IS abundance in bacterial genomes

Early studies of IS in bacterial genomes investigated their sequence similarity to explore relationships between IS and individuals in a population or across species. Multiple studies found that within *E. coli*, sequences of the same IS in the genome were almost identical, with very few substitutions compared to the rest of the genome, suggesting that the IS had recently entered the genome and then proliferated[73,74]. Within the genus *Escherichia*, both *E. coli* and *Escherichia fergusonii* have multiple copies of IS*3*, and the IS*3* sequences within each species

are almost identical to each other. But when comparing IS*3* sequences across species, the sequences were found to be significantly different to one another[73]. This suggests that IS*3* was present in the ancestor of *E. coli* and *E. fergusonnii*, and little transfer of IS*3* has occurred between them since that divergence[73]. A similar phenomenon is observed for IS*200* between *E. coli* and *Salmonella*[74]. However, some IS do cross species boundaries. In *Escherichia*, IS*1* sequences are almost identical, even across species[73]. This suggests that IS*1* has been horizontally transferred into each *Escherichia* species since their divergence[73,75].

Differences in IS insertion sites of closely related strains are sometimes used to type bacteria of the same species. In *E. coli*, six different IS - IS*1*, IS*2*, IS*3*, IS*4*, IS*5* and IS*30*, were observed across 71 strains[76]. The copy numbers of each IS were sufficient to distinguish between closely related strains[76]. Other bacterial species where IS typing has been applied include *Salmonella* (IS*200*)[77], *Shigella* (IS*1*)[78], *M. tuberculosis* (IS*6110*)[79,80], and *Yersinia* (IS*100*)[81]. More recently, some have suggested using IS*16* as marker for *Enterococcus* in epidemiological settings[82]. However, understanding how IS copy number changes in a genome is important if IS locations are going to be used as markers, as the rate of change of the marker influences inferences about how long a genome has been evolving[83].

Many studies have hypothesised about the causes for IS abundance in some bacterial genomes, but only a few studies have attempted to tease apart the complicated dynamics of IS. Two studies have examined IS in hundreds of bacterial genomes, across a wide variety of phylogenetic groups[84,85]. They explored associations between IS copy number and IS dynamics, including the rate of HGT, genome size and pathogenicity, as well as investigating whether IS dynamics follow a neutral model of evolution or a model that includes purifying selection[84,85]. Firstly, neither study found evidence of phylogenetic specificity for many IS families. IS*1*, for example, is often found in the Enterobacteriaceae, but can also be found in phylogenetically distant groups of bacteria[84]. HGT is important for the introduction of IS into a bacterial genome, but is not required for maintaining copy number[84,85]. In these studies, the pathogenicity of a species was not associated with IS copy number[84]. Rather, larger genomes were associated with higher IS copy numbers, though it has been speculated that this may be due to the fact that bacteria with larger genomes occupy a wider variety of ecological niches, so require the additional genome plasticity that IS provide[84,85]. In general, IS copy numbers remain stable across a wide variety of bacterial species, with only a few species such as *Y. pestis*, *Shigella*, *Salmonella* and *A. baumannii* currently undergoing transient expansions of some IS

families. Despite the burst of transposition within these species, it is likely that over longer evolutionary time periods, these additional IS will gradually be eliminated from the population[86].

1.2.2 The effect of IS on bacterial evolution

Early studies of IS investigated the effect they had on the evolution of bacteria in an experimental setting[87,88]. These studies examined the role of IS in adaptation to the environment, structural changes IS created in the genome, and how the rate of IS transposition could be affected by different environmental conditions.

1.2.2.1 Adaptation through IS-mediated structural rearrangements

IS are important for the fast evolution of new traits that assist bacteria in surviving under environmentally stressful conditions. Naas *et al.*[87] studied 30 year old stab cultures of *E. coli* and found that IS-mediated genome rearrangements and deletions were common causes of mutation within these populations, and these mutations aided survival in the nutrient limited conditions of this storage method. Another study followed an *E. coli* population over 10,000 generations[88]. Over the course of the experiment, large scale genome rearrangements and deletions were common. These larger, structural genome changes were often IS mediated, and were more important for the adaptation of the population than point mutations, creating diversity within the population[88]. An extension of this study observed the dynamics of IS over 20,000 generations of *E. coli*, again finding that IS were crucial for creating variation under stressful environmental conditions, as they regularly created beneficial mutations that would have been impossible to achieve with simple point mutations[89]. Each of these studies points to the influence of IS for evolving new traits within bacterial populations, allowing bacteria to maintain highly plastic genomes and quickly generate variation.

1.2.2.2 Rate of transposition in experimental settings

Experimental studies have investigated the transposition rate of IS. IS were frequently found to have a rate of gain an order of magnitude greater than their rate of loss, with rate estimates ranging from 10^{-3} to 10^{-7} per element per generation, depending on the IS[5,76]. There is a

significant relationship between copy number of an IS and the number of transposition events (either gain or loss) observed in a population[90]. However, these rates differ depending on the IS in question. For example, for IS*1* within *E. coli*, excision is rarer than gain[90]. Rates of excision and gain for IS*2* and IS*5*, however, are similar[90]. IS*1*, IS*150* and IS*5* in *E. coli* were found to change copy number rapidly, and contributed the most to structural change compared to other types of mutation[76,87,88]. Different transposition mechanisms of these IS, and how they are regulated (see sections 1.2.4 and 1.2.4.4), are likely to be the cause of variation of transposition rates between different IS.

1.2.2.3 IS-mediated host adaptation

IS can influence bacterial evolution by inactivating genes. This is usually achieved through deletions of segments of the genome and the formation of pseudogenes. For example, IS played a crucial role in the evolution of *Mycobacterium leprae*, the agent of leprosy. Within *M. leprae*, 50% of the genome is entirely pseudogenes, many of these inactivated by IS[91]. *Bordetella pertussis*, the agent of whooping cough, evolved from its non host-restricted relative, *Bordotella bronchiseptica* via multiple different IS-mediated mechanisms[92]. For example, IS*481* mediated large deletions via homologous recombination, and formed multiple psuedogenes[92]. A similar effect was observed in *Y. pestis*, which evolved from *Y. pseudotuberculosis* through IS-mediated genome decay[60]. In *Burkholderia mallei*, the causative agent of equine disease glanders, IS-mediated recombination events, genome rearrangement and pseudogene formation have aided its evolution from the opportunistic human pathogen *Burkholderia pseudomallei*[53,93]. Within *E. coli* O157, IS have interrupted phage or prophage-like regions, suggesting that IS are important for the inactivation of these elements[94]. In addition to phage, there were interruptions in some virulence associated genes, such as curlin biosynthesis, indicating that IS may be responsible for altering the virulence phenotype of these strains[94]. Formation of pseudogenes has also played a role in the evolution of *Shigella* from *E. coli*, where different IS have been responsible for the inactivation of genes that inhibit the ability for *Shigella* to cause disease in humans[95] (discussed in section 1.4.2). In addition to increasing virulence, IS have also been shown to be important factors for combating host defenses by altering surface antigens[59,92].

1.2.2.4 Mobilisation of genes via IS

IS are able to mobilise genes and place them in new genetic contexts through the creation of composite transposons. In a composite transposon, two IS flanking a region of DNA can coordinate their excision, taking the DNA between them and inserting the entire unit elsewhere. If a composite transposon mobilises into a plasmid, it provides further opportunities for the transposon to be transported into novel genetic contexts in different host cells. Transposons have been implicated in the spread of important bacterial toxins, such as the heat stable ST toxin in *E. coli*, which is mobilised by two copies of IS*1* in the transposon Tn*1681*[96,97]. There are several examples of transposons mobilising antibiotic resistance genes into new contexts. One of the first examples of IS mobilising antibiotic resistance genes is Tn*9*. Tn*9* is flanked by copies of IS*1*, and carries the resistance gene *cat*, which confers resistance to chloramphenicol[98]. Other examples include the dissemination of the metallo-beta-lactamase gene, $bla_{NDM\text{-}1}$, within the *Enterobacteriaciae* via Tn*3000* that has caused increased resistance to carbapenem antibiotics[99]. In *A. baumannii*, two copies of ISAba125 have flanked an aminoglycoside resistance gene, *aphA6*, forming the transposon Tn*aphA6* and spreading aminoglycoside resistance[100].

The movement of antibiotic resistance genes is compounded when multiple transposons move into the same backbone, allowing several antibiotic resistance genes to move as a unit. For example, in *S.* Paratyphi and *S.* Typhi, Tn*21* has inserted into Tn*9*, and then Tn*6029* has inserted within Tn*21*, creating a large transposon encoding several antibiotic resistance genes[101]. The entire unit is mobilised by the two copies of IS*1* that make up Tn*9*[101]. This compound transposon is commonly found in IncHI1 plasmid backbones, and so this plasmid now carries multiple antibiotic resistance genes[52]. IncA/C_2 plasmids carrying multiple carbapenem resistance genes have been found in various Enterobacteriaceae[102]. The mobilisation of the carbapenem resistance gene bla_{KPC} within Enterobacteriaceae has been aided by the highly mobile nature of the transposon Tn*4401*[103]. In this example, Tn*4401* has mobilised into multiple plasmid backbones, aiding its dissemination in a hospital environment[103]. In addition to plasmids, antibiotic resistance genes within transposons can accumulate in genomic islands, which can then also move as an entire unit. Examples of genomic islands carrying large numbers of antibiotic resistance genes include AbaR, the *A. baumannii* resistance island in *A. baumannii*, the SGI in *S. enterica*, and the SRL in *Shigella*[33,38,104].

1.2.2.5 IS-mediated upregulation of gene expression

Gene expression can be altered by IS through the creation of transient promoters after insertion of an IS near a coding region of DNA. Many IS contain a -35 promoter region within their flanking IR. If an IS inserts an appropriate distance from a -10 promotor region in the genome, this can create a new promoter that is able to regulate expression of the gene downstream of the IS[9]. In these situations, genes downstream of this new promoter become constitutively expressed. IS-mediated gene upregulation has been linked to increased virulence in various pathogens. For example, in *M. tuberculosis*, IS*6110* has been found to mediate gene upregulation during growth within monocytes[105]. IS*6110* has also been implicated in increased virulence of *Mycobacterium bovis*, where insertion upstream of the *phoP* virulence gene has been linked to an outbreak of the *M. bovis* B strain in Spain[106]. Within *Neisseria meningitidis*, the insertion of IS*1301* into the capsule locus increases expression of the capsule genes, enhancing capsule biosynthesis and generating a more virulent phenotype in human hosts[107].

Bacterial chromosomes are not the only DNA molecules to benefit from modified gene expression caused by IS. Plasmids can also be activated and given a broader host range through IS upregulation. For example, an ISPst4 insertion upstream of the *oriV* gene in plasmid pUC has allowed this plasmid to be more successful at surviving inside cells of the pathogen *Pseudomonas stuzeri*[108].

IS-mediated gene upregulation can also give bacteria the ability to adapt to new environmental niches by altering their metabolic pathways. An IS*3* located upstream of the threonine operon in *E. coli* enables the bacterium to use threonine as its sole carbon source[109]. IS insertions have also been shown to activate cryptic genes, such as an IS upstream of the *bgl* operon within *E. coli*, activating this region and enabling the bacterium to use arbutin as its sole sugar source[110]. Resistance to environmental compounds, such as bromoacetate, can be acquired by up-regulation of a bromoacetate resistance gene in *Xanthobacter*[111].

1.2.3 IS and the development of antimicrobial resistance in bacteria

In recent years, bacteria have become increasingly resistant to multiple antibiotics and resistance is becoming a major concern. Antimicrobial resistance (AMR) can occur through a

variety of mechanisms, including inactivation of the antibiotic[112], creating a new target for the antibiotic[113], pumping the antibiotic out of the cell[114] or mutation of the target of the antibiotic[115]. These mechanisms can evolve via various genetic mutations, including acquiring genes that will degrade or create new targets for an antibiotic, single point mutations in genes targeted by antibiotics so they no longer bind, and acquiring efflux pump genes to pump antibiotics out of the cell[116]. Some bacteria already contain efflux pumps making them intrinsically resistant to certain antibiotics[117,118]. Mutations that confer increased expression of an intrinsic or acquired efflux pump systems have been shown to increase resistance to multiple drugs, and antibiotics of almost every class are able to be neutralised by the effect of efflux pumps[113,119]. AMR in human pathogens has arisen primarily through the importation of resistance genes via HGT[120,121]. Often resistance genes are clustered together in smaller mobile elements such as transposons or integrons, which can then be transferred into plasmids that provide additional mobility between cells (section 1.1.1.6)[13,122–124].

IS are often associated with resistance, as they are able to both mobilise genes and alter gene expression (section 1.2.2.4 and 1.2.2.5). For example, several plasmids have recently been shown to carry the Tn*3000* transposon, flanked by two IS*3000*'s, which has captured a bla_{NDM-1} gene that confers resistance to carbapenems[99]. Additionally, IS-mediated gene interruptions can confer resistance by preventing gene expression. This has been observed in *K. pneumoniae*, where various IS have interrupted the regulator *mgrB*, conferring resistance to colistin[125].

Some IS, such as IS*1* or ISAba1, are able to alter gene expression through their strong promoters[126,127]. If these IS insert themselves upstream of a gene they are able to increase the expression of that gene. For example, IS upstream of an efflux pump promoter can cause these genes to be either constitutively expressed, or expressed at higher levels, producing resistance to additional antibiotics. This has been demonstrated in multiple bacterial contexts, including fluoroquinolone resistance in both *E. coli* and *A. baumannii*[128,129]. Resistance to third generation cephalosporins in *A. baumannii* can be caused by the insertion of either ISAba1 or ISAba125 upstream of the beta-lactamse *ampC*[130,131]. This entire unit has been mobilised by flanking IS and transferred to other bacteria[131]. Upregulation of bla_{OXA-57} by IS*18* has been described in *Acinetobacter bereziniae* and *A. baumannii*, generating antibiotic resistance to recent drugs, such as the carbapenems meropenem and imipenem[132].

1.2.4 IS transposition mechanisms

IS encode several different types of transposases (DDE and DEDD, section 1.2.4.1, HUH, section 1.2.4.2 and serine transposases, section 1.2.4.3), all of which have different chemistries that catalyse their mobilisation reactions. Each transposase type is associated with one or more IS families (section 1.2.5). Within these different transposase types, there is additional complexity surrounding the exact transposition reaction that takes place, including whether a copy of the IS is left behind during transposition, or if the IS excises completely.

1.2.4.1 DDE and DEDD transposases

The majority of known IS families currently described use DDE transposases, where the active site for the transposase enzyme is an amino acid triplet; Asp (D), Asp and Glu (E), (DDE)[72]. IS with DDE transposases have IR at both ends that the transposase use as binding targets. The IR has two domains - one for binding the transposase, and one for cleavage and strand transfer[9]. DDE transposases use either one or two Mg^2+ cations to catalyse DNA cleavage. Different spacings of the DDE motif are found, with members of the same IS family having similar DDE spacings. DDE transposition will often create DR, also known as target site duplications (TSD), of the sequence that they transpose into. This is usually because cleavage of the DNA by the transposase produces two complementary single stranded ends that need to be repaired by the host cell, resulting in small DR flanking the IS element[9].

There are three possible transposition mechanisms used by DDE transposases, and these depend on the specific IS family. In one mechanism, known as copy-out paste-in, a single strand at one end of the IS is cleaved by the transposase, which then attaches itself to the other end of the IS on the same strand, circularising itself (Figure 1.2a). The IS then replicates this circle to create a circularised double stranded DNA intermediate, and repairs the excision (Figure 1.2a). The circularised intermediate can then un-circularise itself and insert into a new DNA molecule, creating an additional copy of the IS. The other two mechanisms are both cut-and-paste methods of transposition. In one, the DDE transposase cleaves both ends of the IS, leaving two 3'-OH ends. The two 3'-OH ends then react with the phosphodiester bond of the target DNA molecule, allowing the IS to insert itself[133,134] (Figure 1.2b). In this reaction, a small 2 bp scar is left behind (green boxes, Figure 1.2b). In the second type of cut-and-paste transposition, both strands of the donor DNA molecule are cleaved, and each strand of the IS reacts with the other, forming a

hairpin intermediate (Figure 1.2c)[72].

Only one IS family, IS*110*, is known to use DEDD tranposases, and not much is currently known about its specific transposition mechanism. They are similar to the four-way Holliday junction resolvase *ruvC*, with the catalytic centre in a similar location[135]. DEDD transposases have similar chemistry to DDE transposases, but do not require terminal IR, and do not generate DR upon insertion[136].

1.2.4.2 HUH transposases

Two IS families, IS*91* and IS*200*/IS*605*, use very different transposases known as HUH. HUH transposases are named after a conserved pair of histidine residues that are separated by a hydrophobic residue (U)[137]. During transposition, a tyrosine residue creates a bond between the donor DNA molecule and the transposition enzyme[137]. IS that use HUH transposases do not have IR or create DR when inserting[138,139]. Instead of IR at the ends of the IS, HUH transposases require hairpin secondary structures to be formed at the ends of the element[137]. There are two major HUH transposase families; Y1 and Y2. Y1 is associated with the IS*200*/IS*605* family, and Y2 with IS*91*[139,140]. The families are divided by the number of Y residues they use as catalytic sites - either one or two, and both families appear to carry out transposition in different ways, though not much is known about the specific mechanisms[141].

1.2.4.3 Serine transposases

A single IS family, IS*607*, uses serine transposases[72]. These transposases are closely related to the Tn*3* family resolvases, and are hypothesised to act in a similar fashion, involving a double stranded circular DNA intermediate[142]. DNA cleavage occurs when an active serine residue reacts with the phosphate DNA backbone, generating a free 3'-OH end[141]. Members of this family are present as incomplete copies within eukaryotic genomes, and are currently the only prokaryotic IS family to have been found in several eukaryotes[143].

1.2.4.4 IS regulate their transposition

IS transposition is regulated via a variety of mechanisms. Many IS are able to express multiple protein products through frameshifting their open reading frames. This can be accomplished

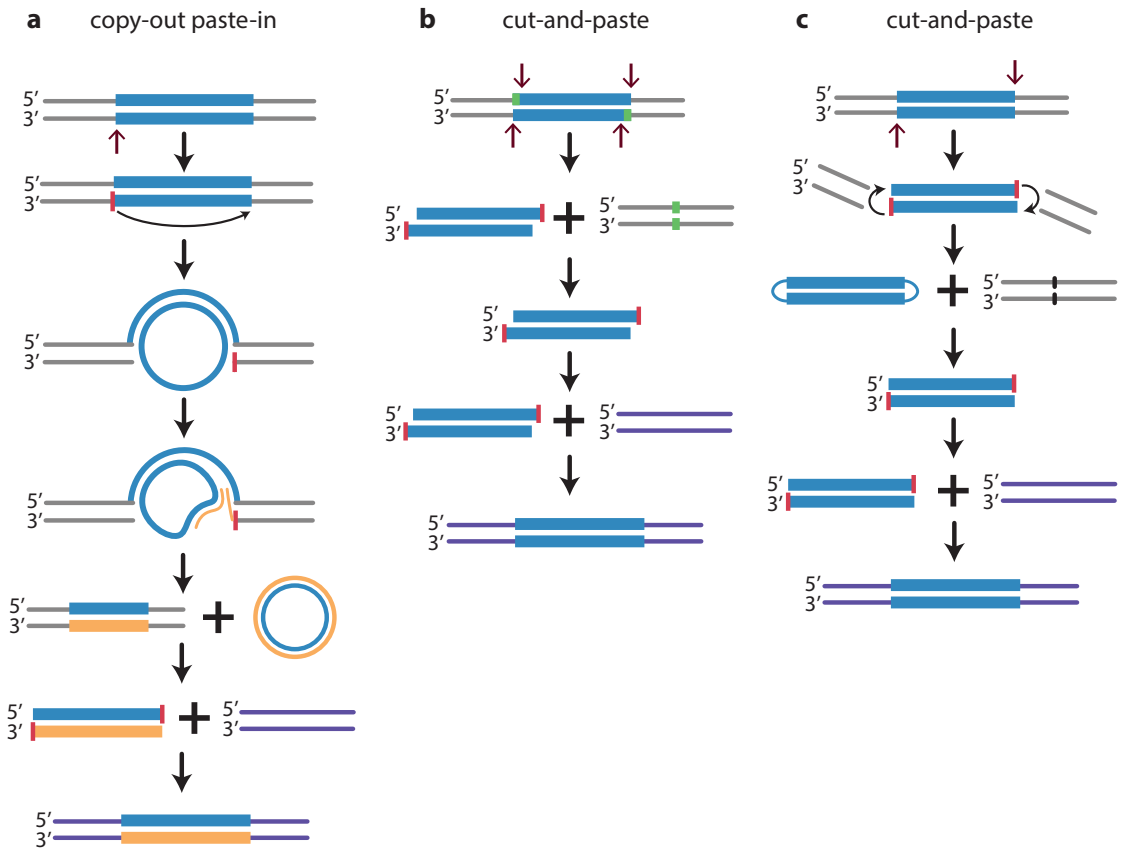

Figure 1.2: Different IS transposition mechanisms. Figure adapted from Siguier *et al.*, 2015.[72]. **a,** copy-out paste-in mechanism. Blue boxes show the IS and the grey lines are the donor DNA molecule. Maroon arrow shows where the transposase cuts. In this transposition mechanism, the bottom strand is cut and then circularises itself, with the red bar indicating where the cut site is. DNA replication then occurs to repair the single strand break, repairing the IS and creating a double stranded circular DNA molecule which can then insert itself into a new DNA molecule (purple lines). **b,** one type of cut-and-paste. Blue boxes show the IS, with maroon arrows indicating the transposase cut sites, some of which occur 2 bp within the insertion sequence (green boxes). The IS is then excised from the donor DNA molecule (grey lines) and can then 'paste' itself into a new DNA molecule (purple lines). **c,** a second cut-and-paste mechanism, with a single cut on each strand (maroon arrows). The IS then forms hairpin structures at either end of the IS molecule, before inserting itself into a new DNA molecule (purple lines).

either by slippage of the RNA polymerase, or by slippage of the ribosome[144]. The rate of frameshifting determines the transposition frequency of the element, as often both protein products are required for successful transposition[9]. Some IS have been shown to have preferential *cis* activity, where the transposase acts preferentially on the IS that it was transcribed from, rather than other copies of the same element within the genome[9]. Transposase activity can be blocked by the use of antisense RNA. Some IS, such as IS*10*, contain promoters within their transposase genes that encode for short antisense RNA molecules, responsible for blocking transcription[145]. Other IS have been shown to be regulated by temperature, for example IS*30* and IS*911*[146,147]. Many IS have target site specificity, which may limit their transposition into other bacterial species that do not have the correct targets[9].

In addition to element-specific regulatory mechanisms, there are several host factors which have been shown to influence transposition. Host factors can act at various stages of transposition, by inhibiting transpososome assembly or by altering the way DNA repair takes place. DNA chaperones IHF, HU, Fis and HNS have all been implicated in the regulating transposition[148]. IHF has been shown to assist with transposition of IS*10* and IS*1*[149,150]. DAM methylation has also been implicated in regulating transposition by either interfering in the actual transposition step or altering the expression of the transposase[9].

1.2.5 Grouping of IS into IS families

IS are grouped into families based on three factors. The first is the type of transposase they encode, as discussed in section 1.2.4. Secondly, some IS contain additional accessory genes that regulate their transposition, and these genes are specific to each IS family[136]. Finally, sequence similarity to other IS influences the organisation of IS into IS families.

The IS families discussed below are those that were detected during the analyses present in this thesis. The majority of these IS are found within *Shigella spp.*, and are discussed in greater detail in Chapters 4 and 5. The final IS family discussed, IS*6*, is found in a variety of bacterial species and is linked to the movement of antibiotic resistance genes. The IS*6* family is discussed in greater detail in Chapter 3.

1.2.5.1 IS*1*

IS*1* was one of the first IS to be discovered, and has been identified in over 40 bacterial and archael species[72]. It uses the DDE transposase system, and has a 23 bp IR. IS*1* generates DR of mainly 9 bp, though DR of 8, 10 or 14 bp have also been reported. IS*1* consists of two open reading frames, *insA* and *insB*, that are expressed from a promoter located in the left IR. Only two products are produced. InsA binds to both IR and regulates transposition, whilst the frameshift product of *insA* and *insB*, InsAB', produces the transposase itself. No product for *insB* alone has been described. During transposition, members of this family can integrate into a new site using one of two methods - cointegrate formation or the copy-out paste-in approach[151]. Transposition rate appears to be controlled by the ratio of InsA/InsAB' product[72]. Members of IS*1* show a preference for AT rich regions within genomes[6]. Frequently, IS*1* members have been found to form transposons in either direct or inverted orientation, often carrying antibiotic resistance or virulence genes (for example Tn*9*, section 1.2.2.4)[96,101].

1.2.5.2 IS*3*

IS*3* is one of the largest IS families, containing 554 members in over 270 different bacterial species, and range in size from 1,200 - 1,550 bp[72]. Elements in the IS*3* family have DDE transposases, with terminal IR, and create DR of 3-4 bp. Their transposase is generated by a fusion product of *orfA* and *orfB*, caused by translational frameshifting in a similar fashion to IS*1*. Members of the IS*3* family use the copy-out paste-in transposition mechanism.[72]. One member of the IS*3* family, IS*911*, has been shown to act in *cis*, with the transposase preferentially targeting the IS element it was expressed from[152]. Some members of this family, such as IS*2* and IS*911*, have been shown to contain strong promoters in the -35 and -10 regions of the IR, allowing them to influence expression of downstream genes[6].

1.2.5.3 IS*4*

The IS*4* family, consisting of over 200 members, has recently been redefined to include seven subfamilies (IS*231*, ISH8, IS*4Sa*, IS*4*, IS*10*, IS*50* and ISPepr1), and three emerging IS families (IS*701*, ISH3 and IS*1634*)[72]. All members of this family use a DDE transposase, but contain a large beta strand between the final D and E residues in the motif[153]. Unlike many other IS

families, no frameshifting occurs to generate the transposase[154]. All members of this family also contain a YREK tetrad motif, hypothesised to be involved in DNA cleavage[153,155]. Members of this family transpose via the cut and paste mechanism, and generate a hairpin intermediate during transposition[6,153]. Several members of this family are involved in the generation of transposons, especially IS*10* and IS*50*, which form the transposons Tn*10* and Tn*5* respectively[156,157].

1.2.5.4 IS*21*

IS*21* elements encode two genes, *istA*, which encodes a DDE transposase, and *istB*, which encodes a helper gene[72]. Both genes are required for successful transposition[158]. The IR of these elements contain several repeats that are thought to be involved in transposase binding[6]. This family creates DR of approximately 4 bp[6]. IS*21* preferentially targets sites where there are other IS*21* elements, generating tandem duplications, and uses a similar copy paste mechanism to other IS families[72].

1.2.5.5 IS*66*

IS*66* contains three open reading frames (orfs) - *tnpA*, *tnpB* and *tnpC*, with a 30 bp IR at each end[72]. IS*66* elements use a DDE transposase that is encoded by *tnpC*, and transposition of this element creates 8 bp DR. Mutations within *tnpA* or *tnpB* have been shown to reduce the level of transposition activity, however some elements that do not contain both *tnpA* and *tnpB* are still able to successfully transpose[72,159]. In addition, some members of this family that contain only *tnpC* still transpose[136]. This family is grouped into three major classes depending on their gene content - one group contains all three genes, *tnpA*, *tnpB* and *tnpC*, another group lacks *tnpA* but contains *tnpB* and *tnpC*, and the final group contains only *tnpC* but commonly has passenger genes downstream of *tnpC*[72]. Little is known about the transposition mechanism or target site preferences of this family.

1.2.5.6 IS*630*

Members of the IS*630* family contain a single open reading frame (orf), but in some cases the orf is spread across two reading frames, so may be frameshifted to produce the transposase[72].

This family uses a DDE transposase, and prefer to insert into a TA dinucleotide[160]. Additionally, members of this family have a striking similarity to eukaryotic transposable elements, especially the Tc*1*, Tc*3* and *mariner* families[161]. When transposing, they appear to have similar transposition mechanisms to their relatives within the eukaryotes, but the exact mechanism is unknown[153].

1.2.5.7 IS*110*

The IS*110* family is the only family known to use a DEDD transposase, with a single orf that encodes the transposase[72]. Due to their transposition method, elements in this family do not have IR, but instead have secondary hairpin structures at their terminal ends. As such, they do not make DR on insertion. During transposition, this family creates circular double stranded intermediates, but it is not known whether these intermediates are caused by a copy-paste mechanism, or an excision mechanism[162,163]. There are several known insertion site preferences for different members of this family. For example, members in the IS*1111* subgroup prefer *attC* integron regions, while others prefer specific DNA sequences, or specifically target the ends of other IS elements[164–166].

1.2.5.8 IS*200*/IS*605*

The IS*200*/IS*605* family use HUH transposases. They are divided into three main groups based on the number of genes they contain[72]. One group contains both *tnpA* and *tnpB*, another only *tnpA*, and the final group only *tnpB*[139]. Experiments have shown that only *tnpA* is required for transposition, so elements that contain only *tnpB* may be unable to transpose[72,167]. Regulation of transposition may be mediated by *tnpB*[72,167]. Members of this family transpose using a peel-and-paste mechanism. One strand of the IS is excised and creates a single stranded circle[168]. This circle then inserts itself into a single stranded target molecule, preferring to insert themselves into the lagging genome strand of a replicating genome, creating an orientation bias[168]. As they use HUH transposases, they create no DR upon insertion.

1.2.5.9 IS*6*

Members of the IS*6* family use DDE transposases. They contain a single orf that is promoted from within the left IR, and create 8 bp DR upon insertion[72]. No target specificity has so far been described. Several members of this family are well known for their role in mobilising antibiotic resistance genes. For example, IS*257* has played an important role moving resistance genes within *S. aureus*.[169] IS*257* also contains a promoter in its left end, so can up-regulate genes that are downstream[169]. Another member of this family, IS*26*, has recently been implicated in antibiotic resistance regions in many bacterial plasmids and genomes, and has been involved in large scale genome rearrangements[170]–[172]. IS*26* is discussed in further detail in Chapter 3.

1.3 Approaches in genomics to investigate bacterial evolution

The ultimate aim of genomics is to analyse bacterial DNA, the raw output of which is short sequences of DNA. These DNA sequences allow us to answer many important biological questions. The following sections detail each of these approaches, their technical challenges, and the different ways they are used to investigate the evolution of bacterial genomes.

1.3.1 Sequencing bacterial genomes with whole genome sequencing

The study of bacterial evolution has recently benefited from the use of high throughput, whole genome DNA sequencing (WGS). Initially, bacterial genomes were sequenced using the Sanger shotgun sequencing method. However, this method was slow, and advances in sequencing technology soon produced faster methods for sequencing whole genomes.

Early next generation sequencing was performed using Roche 454 or IonTorrent sequencing machines. Roche 454 sequencing involves the fragmentation of input DNA into small pieces, which are then attached to beads. PCR amplifies each small fragment on the bead, and the beads are then affixed to a glass slide. A single nucleotide type (either A, C, G or T) is washed over the slide, and if incorporated into the new DNA strand, a bioluminescent signal is produced and recorded. If multiple nucelotides are incorporated (eg, a string of A's), then the strength of the emitted light is increased. The amount of emitted light was analysed to determine how many bases had been incorporated. IonTorrent sequencing works in a similar manner. After the

DNA is fragmented, a single nucleotide type is washed over the wells. Instead of incorporated nucelotides emmitting a bioluminescent signal, incorporated nucleotides emit a single H+ ion. The release of this H+ ion generates a small change in pH, which is detected by the sequencer. Addition of several of the same nucleotide causes a larger pH change.

Both Roche 454 and IonTorrent sequencing are high throughput, and produce reads between 400 - 700 bp in length. However, both platforms suffer from high rates of indel errors, and errors within homopolymer regions[173]. These limitations have made these sequencing methods less popular than alternate short read sequencing methods, such as Illumina.

The Illumina platform also produces short reads, with DNA fragmented into small segments and affixed to wells within a flow cell. Unlike 454 and IonTorrent, all four nucleotides are washed over the wells, and incorporation of a nucleotide produces a fluorescent signal, with a different colour emitted for each nucleotide type. A photo is taken of the flow cell to determine which nucleotides have incorporated into which fragments. As Illumina uses imaging to determine which base has been incorporated, Illumina reads do not suffer from the same homopolymer errors as 454 and IonTorrent. Recent advances in Illumina chemistry have improved read length from ~50 bp, in the beginning, to ~300 bp on the Illumina MiSeq today. In addition to their high accuracy, Illumina sequencing also produces paired end short reads - these are sequences of DNA taken from the same DNA fragment, but separated by a certain number of base pairs, called an insert size. This additional information is frequently leveraged by downstream analysis approaches.

In addition to the above short read sequencing methods, more advanced sequencing methods that produce significantly longer reads than Illumina, IonTorrent, or Roche 454, have been developed. The two main contenders in this arena are PacBio and Oxford Nanopore. Both methods still require the fragmentation of DNA, but these methods use much larger DNA fragments. In PacBio sequencing, a single stranded DNA fragment is placed inside a well containing a single DNA polymerase. Fluorescent tagged nucleotides are continually fed into the wells, and as each nucleotide is incorporated it emits a coloured signal, which is captured by a video camera. Oxford Nanopore uses a charged membrane containing many pores. As the DNA fragment is pulled through the pore, changes in the electrical current flowing through the membrane are measured. The electrical current alters depending on which nucelotides are sitting in the pore at any one time. A computer algorithm translates these current changes into bases. In both methods, the accuracy of the reads produced is much

lower than those produced by Illumina. However, their large size (from 10 kbp to 200 kbp) allows the user more easily reconstruct the complete genome. Overall, short read sequencing using Illumina is generally the cheapest, and currently in 2017 is the most common sequencing approach. Long read sequencing with PacBio is more expensive, costing ~10 times more per bacterial genome than Illumina (~$1,000 vs ~$100) to generate sufficient sequence data. Long read sequencing with Oxford Nanopore is becoming much more affordable, with multiplex sequencing protocols bringing the cost per bacterial genome down to the same level as Illumina. However, Oxford Nanopore technology is still undergoing extensive development and is currently less stable and accurate than Illumina or PacBio technology; indeed Oxford Nanopore base call accuracy is not yet sufficient to produce accurate consensus sequence, and Illumina short reads are still required to create an error-free genome assembly[174].

This thesis is primarily concerned with short reads produced by the Illumina sequencing method, and the following sections discuss the strengths and limitations of using short Illumina reads in genomics.

1.3.2 Reconstructing bacterial genomes with assembly

Assembly aims to reconstruct the genome back into its original state, before it was fragmented for sequencing. The most common method of assembly today is the de Brujin method, first reported in 2001 by Pevzner *et al.*[175]. In de Bruijn assembly, reads are split into smaller sequences, called kmers[176]. Kmers are compared to determine which ones overlap with others, forming a graph of longer, contiguous sequences (contigs), and their connections to other contigs[176]. This form of assembly is performed *de novo*, without the use of a reference genome to guide the assembler. As the reads are short, they are frequently unable to completely span regions of the genome that are longer than the read length and repeated many times[177]. These repetitive regions are difficult to assemble, as they create multiple paths through the assembly graph that the assembler is unable to resolve[177]. Due to the difficulty in resolving these regions, many contigs will be reported for a single genome assembly[177]. Taken together, these contigs usually represent the majority of the genome, but they will not be connected in the correct order. Assembly is also a computationally intensive process, often taking many hours and several gigabytes of memory to assemble a single genome.

Despite the limitations of assembly, assemblies are useful for investigating the evolution of

bacterial genomes. As assemblers attempt to reconstruct the original DNA sequence, they are good for identifying novel DNA sequences or genes that have been gained by the genome, either in the form of genomic islands, plasmids, or other smaller mobile elements. Genome assemblies may be interrogated to detect large structural rearrangements within the genome. Smaller variants, such as SNPs and indels, can be detected using assemblies, but assemblers can introduce single base errors into assemblies and so there can be a lack sensitivity for detecting this type of variation. As IS are frequently present in multiple copies throughout a genome, they can be the repetitive regions responsible for breaking the assembly into multiple contigs. Genomes with a high IS burden will be more difficult to assemble than those without. Variation arising from complex structures such as antibiotic resistance regions, that contain many IS elements, can therefore be difficult to resolve using assemblies.

1.3.3 Detecting variation in bacterial genomes using mapping

Mapping is the process of taking reads and aligning them to a reference, comparing the reads to a known sequence, to identify differences in the DNA sequence against the reference. This process is very fast and not computationally intensive. Once reads are aligned to a reference, two important metrics can be calculated - depth, which is the number of reads aligning to a particular sequence, and coverage, which is the number of reads spanning a particular sequence. Good depth and coverage of reads mapped to a reference allows the inference of variations between the reads and the reference sequence, and which sequences in the reference are present or absent in the read set.

Mapping approaches are most often used to detect variants arising from point mutations, including SNPs and indels. Previous studies have used SNP data to identify point mutations that are important for adaptation to a new niche. Examples include the loss of fimbrial genes during host adaptation in *S.* Gallinarum or *S.* Typhi[51,61], or the generation of antibiotic resistance against drugs such as fluoroquinolones in *S.* Typhi[178,179]. Mapping can also be used to perform sequence typing, either by detecting allelic variation at loci or determining the presence or absence of specific genes. For example, determining the multi-locus sequence type (MLST) of genomes is frequently done using a mapping approach. In MLST, six or seven loci, which are conserved housekeeping genes, are selected for each bacterial species[180]. At each locus, an identifying number is given to each allele, and the combination of allele numbers results in a sequence type (ST)[180]. MLST is a useful method for grouping bacterial genomes of

the same species without relying on phenotypic characteristics that can easily be horizontally transferred[181]. In addition to determining ST, presence/absence of genes can be quickly inferred using mapping, including antibiotic resistance genes (using databases such as ResFinder[182], CARD[183], or ARGAnnot[184]), virulence genes (using VFDB[185]), or plasmid replicons (using PlasmidFinder[186]).

However, as mapping approaches are always performed using a reference, they are unable to identify novel DNA sequences that are not present in the reference. This makes it more difficult to identify how genomes have evolved by gaining additional genes, especially if those genes are novel sequences. Additionally, it is difficult to determine the structure of the genome if the structure of interest is not present in the reference.

1.3.4 Phylogenetics: understanding evolutionary relationships between genomes

Originally, the relatedness of bacterial isolates was investigated by aligning sequences of single genes and constructing phylogenetic trees to examine their relationships. However, in cases where bacterial genomes are very closely related, such as *Y. pestis*, *M. tuberculosis* or *S.* Typhi, there is frequently not enough variation found in a single gene to resolve relationships between isolates. Since the advent of WGS, short reads can be mapped to reference genomes to generate an alignment of SNPs, and these can be used to create whole genome phylogenies of bacteria. Phylogenies have shaped our understanding of the evolution and spread of many important bacterial pathogens, such as *E. coli*[187], *K. pneumoniae*[188], *Salmonella* Typhimurium[189], *S.* Typhi[190] and *M. tuberculosis*[191]. First demonstrated in methicillin-resistant *S. aureus*, phylogenies can also be used to delineate between outbreaks and transmission events of pathogens[192].

In phylogenetics, the SNP alignment, plus a model of how DNA substitutions occur, combine to construct an evolutionary history. Assessing the number of substitutions per site in a phylogenetic framework allows for the calculation of a molecular clock. The term 'molecular clock' was first introduced by Zuckerkandl and Pauling in 1962[193], who discovered that amino acid changes in hemoglobin appeared to be constant over time, and so the evolution of these sequences were thus 'clock-like'. This method was then extended to DNA substitutions.

In the past, mutation rates were studied experimentally in *E. coli*[194]. Several different *E. coli lacZ* mutants were grown on media containing lactose. The number of generations it took to revert the inactive *lacZ* allele to a working *lacZ* allele was used to calculate the number of substitutions per generation[194]. Mutation rates are now routinely calculated using SNPs from WGS[195]. The rates calculated using WGS differ from those previously calculated using an experimental approach, as these mutation rates are expressed in the number of substitutions per site, rather than the number of substitutions per generation[71]. Evaluation of mutation rates in different bacterial species has shown that evolutionary rate is flexible, and varies across species[196].

Using the molecular clock framework in combination with sampling dates for isolates, SNP data can also be used to perform molecular dating on a phylogenetic tree. First demonstrated in human immunodeficiency virus (HIV), branch lengths on the phylogeny were calibrated using the sampling times of each isolate, to reveal that HIV had likely emerged from Central Africa in the 1960s, before spreading to the Americas in the 1970s[197]. Early methods of molecular dating assumed a constant rate of evolution, or a strict molecular clock. However, it has since been revealed that not all organisms evolve at a constant rate, and so different clock models were proposed that allowed the mutation rate to vary across the branches of the phylogeny[198]. Clock models for dating phylogenies are usually applied using one of two major frameworks - maximum likelihood or Bayesian. Under a maximum likelihood approach, a strict clock must be assumed[199]. However, Bayesian approaches allow for different clock models, such as uncorrelated relaxed clocks[200]. Combining dating approaches with epidemiological data provides an excellent framework for exploring transmission patterns of pathogens[201].

1.3.5 Using genomics to investigate the impact of IS on bacterial genomes

Modern genomics analyses currently investigate hundreds, or thousands, of bacterial genomes at once using the above approaches. This analysis task is a large undertaking due to the size of the bacterial datasets. Currently, many approaches investigating large numbers of bacterial genomes focus on variation arising from SNPs or gene presence/absence, however few studies have attempted to determine the impact of IS on the evolution of bacteria, especially pathogens. Currently, there are few tools available to interrogate IS content in large numbers of bacterial genomes.

A major focus of this thesis is to address this methodological challenge, and then apply this new methodology to novel contexts. *Shigella* is an ideal candidate for addressing the impact of IS on the evolution of bacterial genomes, as *Shigella* genomes harbour hundreds of IS.

1.4 *Shigella spp.*

Shigella is a Gram-negative, rod shaped, intracellular bacterial pathogen that causes diarrhea. It is transmitted via the fecal-oral route through contaminated food or water. *Shigella* consists of four species: *Shigella flexneri*, *Shigella sonnei*, *Shigella dysenteriae* and *Shigella boydii*. *S. flexneri* and *S. sonnei* are endemic and contribute the most to disease burden[202]. In contrast, *S. dysenteriae* is mostly associated with outbreaks, causing severe and life threatening illness[203]. Little is known about *S. boydii*, which is primarily found on the Indian subcontinent[202]. *Shigella* has a very low infectious dose of less than ten bacterial cells[204]. This is likely due to its intrinsic acid resistance, allowing it to survive in the gut, and its ability to down-regulate the antimicrobial peptides intestinal cells excrete[205,206].

1.4.1 *Shigella* pathogenesis

Much of the experimental work on pathogenesis in *Shigella* has been performed using *S. flexneri*. After making their way to the intestine, *Shigella* invade specialised epithelial cells called M cells, where they are transcytosed through the cell and into a waiting macrophage[207]. The macrophage then phagocytoses the bacterial cell, only to have the bacterium escape phagocytosis and induce cell death in the macrophage via the caspase-1 pathway[208,209]. After killing the macrophage, the bacterium then invades a nearby epithelial cell, from the basolateral side of the intestine, and replicates itself within the cytoplasm of the host cell[210]. *Shigella* then hijacks the cell's actin pathway, and uses this to spread into neighboring epithelial cells, bypassing interaction with the host immune system[211].

Throughout this process, several host inflammatory responses occur. Interleukin is produced after macrophage killing, and innate immune cells are recruited to the area. These cells continue to produce interleukins and other host immune factors. All of these host responses contribute to the instability of the intestine, resulting in the watery diarrhea that is a hallmark symptom of *Shigella* infection[212]. In the case of *S. dysenteriae*, this diarrhea is more severe due

to the production of the Shiga toxin. Shiga toxin is cytotoxic, and causes lesions on the colon, kidneys and central nervous system, inducing the more fatal disease type associated with *S. dysenteriae*[213].

1.4.2 *Shigella* has evolved from *E. coli*

The initial discovery of *Shigella* and its clinical disease type led to its original categorisation as a separate genus from the *E. coli* group[214]. *Shigella spp.* were also split into several serotypes based on their O antigen. *S. flexneri* has 14 serotypes, *S. dysenteriae* 15 serotypes, *S. boydii* 20 serotypes, and *S. sonnei* consists of a single serotype[215]. Early studies of *E. coli* and *Shigella* enzymes and sequences revealed a close relationship between these two species[216–219]. Later genomic analysis and phylogenetics revealed that *Shigella* belonged within the *E. coli* group, and that each *Shigella* species was the result of convergent evolution of different *E. coli* lineages[220,221]. Phylogenetic analysis of the regions *thrB-thrC*, *trpB-trpC*, *purM-purN* and *mdh-argR* of both *Shigella* and *E. coli* revealed three main clusters of *Shigella* - C1, C2 and C3[221]. The first cluster was additionally split into three subclusters - SC1, SC2 and SC3[222]. *S. flexneri*, *S. dysenteriae* and *S. boydii* are distributed across the three main clusters, with *S. sonnei*, *S. dysenteriae* serotypes 1, 8 and 10, and *S. boydii* serotype 13 falling as outliers to the three clusters[222,223].

Several evolutionary events have shaped this invasive bacterial pathogen (Figure 1.3). Firstly, each *Shigella* lineage acquired the virulence plasmid pINV. This plasmid contains many of the genes required for invasive infection and survival inside the host cell[224]. The most crucial part of this plasmid is the set of genes known as the *mxi-spa* locus, which encodes a type three secretion system (T3SS). Virulence effector proteins that interfere with host cell processes are translocated from the bacterial cell into the host cell via the T3SS[225,226].

There are two main types of virulence plasmid in *Shigella*, pINV-A and pINV-B. Both types are spread across the different species, and are of the same Inc type, so cannot exist in the same cell together[227]. Strains within C1 carry only pINV-A, and strains in C3 carry only pINV-B[223]. Strains that are outside of the three main clusters can carry either pINV-A or pINV-B. In the case of *S. dysenteriae* 1, this species has a mixed form of the virulence plasmid made up of segments of both pINV-A and pINV-B[223]. Taken together, these results suggest that the virulence plasmid has coevolved with each *Shigella* lineage[222,227,228].

Additional gene gains came from the acquisition of several pathogenicity islands into the

chromosome, all of which are important for virulence[229] (Figure 1.3). SHI-1 encodes several toxins[230], and SHI-2 carries the genes for biosynthesis of the siderophore aerobactin, bacteriocidal genes, and genes that assist with evasion of the host immune system[231,232]. Both SHI-1 and SHI-2 are found in all *Shigella spp*. SHI-3 is found only in *S. boydii*, and is almost identical to SHI-2, but encodes an additional set of siderophore genes[233,234]. SHI-O, found in *S. flexneri* 2a, has the ability to modify the O-antigen on the outside of the cell, enabling *Shigella* to switch serotypes and evade the host immune system[235]. Finally, the acquisition of the SRL, which carries several antibiotic resistance genes, has given some strains across all *Shigella* species resistance to several first line antibiotics, including streptomycin, ampicillin, chloramphenicol and tetracycline[40,236,237]. All of these pathogenicity islands are associated with phage integrases, suggesting that phage have played an important role in *Shigella* evolution[229,235,238].

In addition to gene gain, the loss of six different pathways have been inactivated in *Shigella* compared to *E. coli* (Figure 1.3). For example, unlike *E. coli*, *Shigella* requires nicotine acid for growth, and there is strong selective pressure to inactivate both *nadA* and *nadB*, which are involved in the synthesis of nicotine acid[239,240]. The loss of these is genes is patho-adaptive, as intermediate products in the nicotine pathway have been shown to decrease *Shigella*'s ability to spread intracellularly[241]. The lysine decarboxylase, *cadA*, has also been inactivated in all *Shigella*[233]. This gene produces cadaverine, which inhibits the enterotoxins *Shigella* produces, and when complemented back into *S. flexneri*, it was found to hinder pathogenicity[242]. Additionally, *Shigella* accumulate spermidine within their cells, unlike *E. coli*, through the loss of *speG*, which converts spermidine into the non-reactive acetylspermidine[243]. The loss of this pathway is hypothesised to assist with the survival of *Shigella* within the stressful environment of macrophages[243].

In addition to the loss of the above pathways, *Shigella* has also lost *ompT*, which encodes an outer membrane protease[233]. When *ompT* is complemented back into *S. flexneri*, it has been shown to interfere with *S. flexneri's* ability to manipulate actin, preventing the spread of *S. flexneri* within epithelial cells[244]. All *Shigella* have also lost *argT*, which encodes an arginine/lysine/orthinine binding protein[245]. The reason for this inactivation is currently unknown, however experimental evidence has shown that the presence of *argT* decreases invasion of *Shigella* into HeLa cells[245]. Finally, all *Shigella* species are non-motile, and have inactivated all their flagella and fimbriae genes, likely to prevent recognition by the host

immune system[246,247].

Many of the gene inactivations found in *Shigella* have been caused by IS, either through IS-mediated interruption or IS-mediated genome rearrangements. In a genomic comparative study of *S. flexneri* 2a strain 2457 and strain 301, strain 2457 was found to contain 372 pseudogenes, 30% of which were a direct result of IS-mediated inactivation[248]. Within *S. flexneri* 2a strain 301, IS were responsible for several genome rearrangements[248].

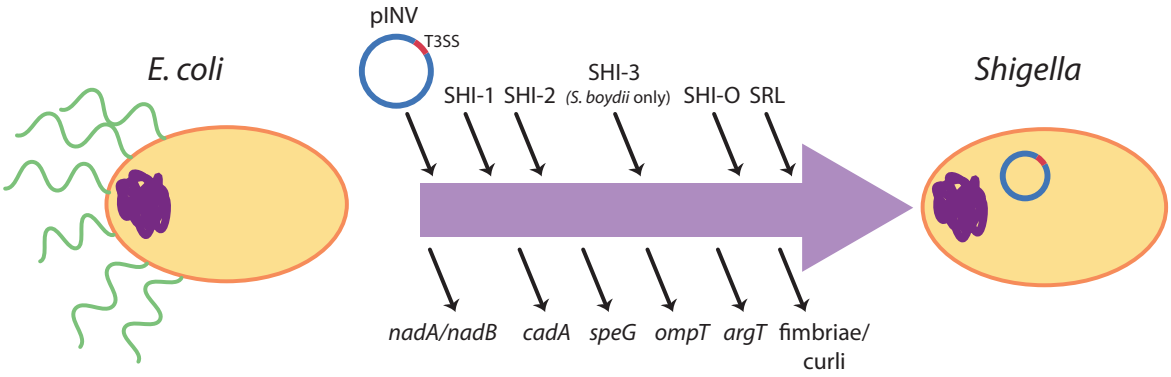

Figure 1.3: Steps in the evolution of *Shigella* from its ancestor, *E. coli*.

1.4.3 Population history and genomic studies of *Shigella spp.*

1.4.3.1 *S. sonnei*

S. sonnei typically causes dysentery in developed countries. Early studies showed that *S. sonnei* is a highly clonal lineage of *E. coli*, with few differences between genomes[249]. However, more recent comparative genomics studies have elucidated the evolutionary history and global population structure of *S. sonnei*, showing that it arose in Europe in the mid to late 17th century and diverged into three main lineages that have each undergone further diversification since the early to mid 19th century (Figure 1.4)[250]. All three lineages originated in Europe, however Lineage III is currently the most prevalent and widespread globally (Figure 1.4)[250]. Recently, *S. sonnei* has been found to be causing increased disease burden in nations undergoing economic development, often replacing previously circulating populations of *S. flexneri*[251,252]. These changes in disease demographics are likely driven by improvements in water quality[251,253]. Most of this replacement is being driven by lineage III, which includes two widely disseminated multi-drug resistant subclades know as Global III and Central Asia III[250,252,254].

1.4.3.2 *S. dysenteriae*

S. dysenteriae was first isolated in Japan by Kiyoshi Shiga, during a dysentery outbreak in the late 1890s that was responsible for tens of thousands of cases and deaths[255]. Since its discovery, this species has commonly been responsible for large outbreaks, including outbreaks in Central America[256], Africa[257] and Asia[258,259] during the twentieth century. As not much is known about the evolutionary history of *S. dysenteriae*, Njamkepo *et al.*. conducted a study of 325 *S. dysenteriae* genomes from 66 countries, spanning the years 1915 to 2011[260]. Isolates from all major outbreaks were included. Phylogenetic analysis of this data showed that the population of *S. dysenteriae* Sd1 had a substitution rate of 8.7×10^{-7} substitutions site^{-1} year^{-1}, and is made up of four lineages, all distributed geographically (Figure 1.5a)[260]. A subset of this data, consisting of 125 spatially and temporally representative genomes, was used to construct a dated phylogeny. It was found that the *S. dysenteriae* Sd1 population arose in ~1747, and spread globally during the late nineteenth century (Figure 1.5b)[260]. During this time, *S. dysenteriae* has slowly acquired antibiotic resistance, primarily through the acquisition of

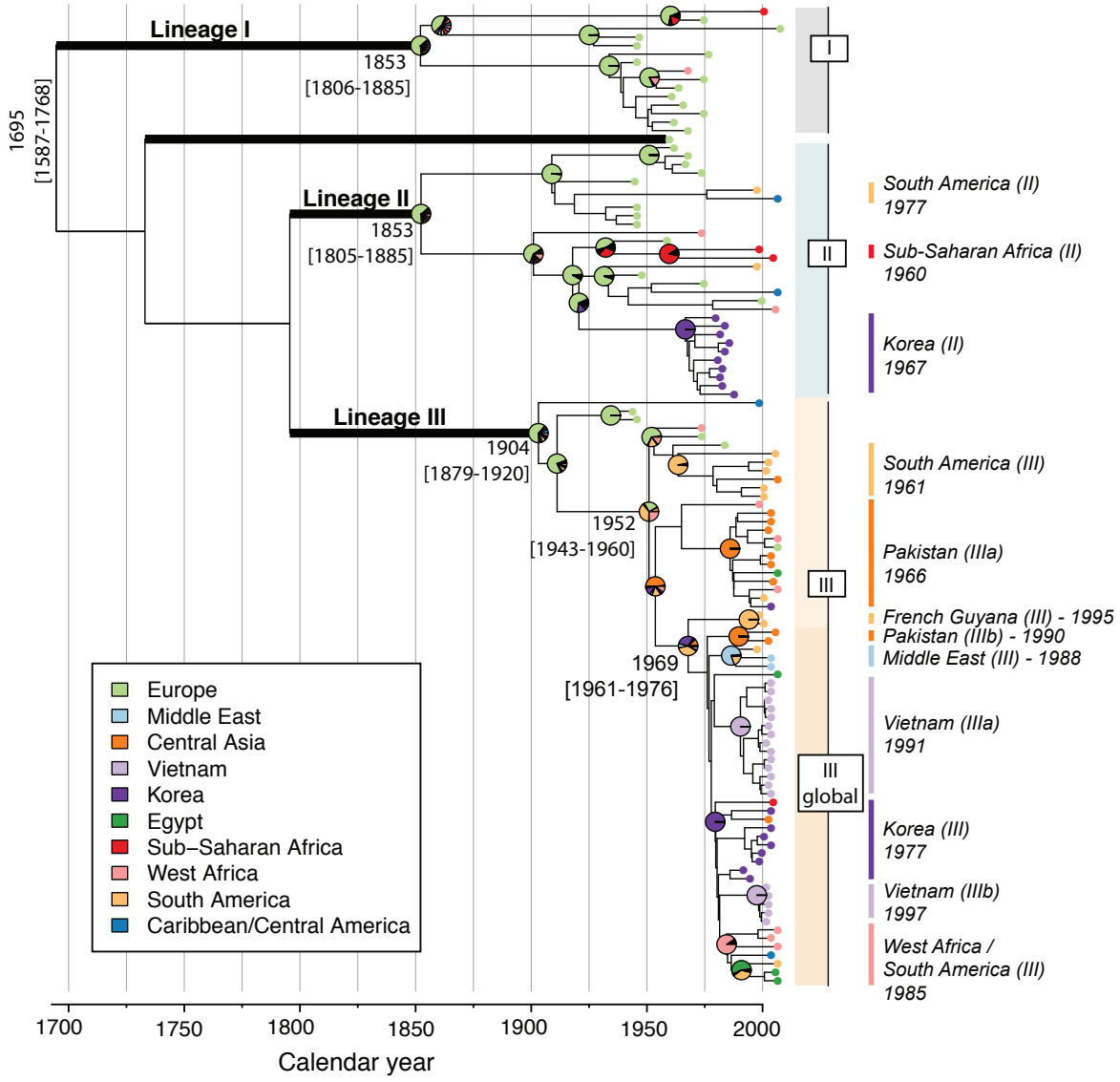

Figure 1.4: Bayesian maximum clade credibility phylogeny for *S. sonnei*. Figure adapted from Holt *et al.*, 2012[250]. Branches defining major lineages are shown in bold (each with 100% posterior support). Pie charts indicate maximum-likelihood estimates for geographic origin of major nodes. Divergence dates (median estimates and 95% HPD) shown for major nodes.

the SRL, large, antibiotic resistance plasmids, and mutations in *gyrA* and *parC*[260].

1.4.3.3 *S. flexneri*

Despite the fact that *S. dysenteriae* is responsible for the majority of dysentery outbreaks, *S. flexneri* is one of the most common *Shigella* species globally, and causes the majority of disease[251,261]. Until recently, little was known about the population structure of *S. flexneri*. Serotyping has traditionally been used to divide *S. flexneri* into different groups, however early phylogenetic studies demonstrated that there are two major groups of *S. flexneri*[262]. One, made up entirely of *S. flexneri* serotype 6, falls within the *S. boydii* cluster[263]. A second group, including all other *S. flexneri* serotypes (1 - 5, X and Y), contains the majority of isolates[263].

To study the population structure of this second group of *S. flexneri*, Connor *et al.*.[264] collected 351 genomes, spanning the years 1914 - 2011. A maximum likelihood phylogeny of these genomes showed that *S. flexneri* consists of seven lineages, all of which are separated by large evolutionary distances, unlike the lineages found in *S. sonnei* and *S. dysenteriae* (Figure 1.6)[264]. All lineages were found in most geographic regions and each consisted of several serotypes (Figure 1.6)[264]. BEAST analysis of each lineage revealed a wide range of lineage emergence dates[264]. The earliest emergence was lineage 4, which arose ~1341, and the most recently emerged lineage was lineage 3, which arose ~1848 (Figure 1.6).

Overall, *S. flexneri* has a very different population structure to *S. sonnei* and *S. dysenteriae*, and is more similar to other pathogenic linages of *E. coli*, such as ETEC.

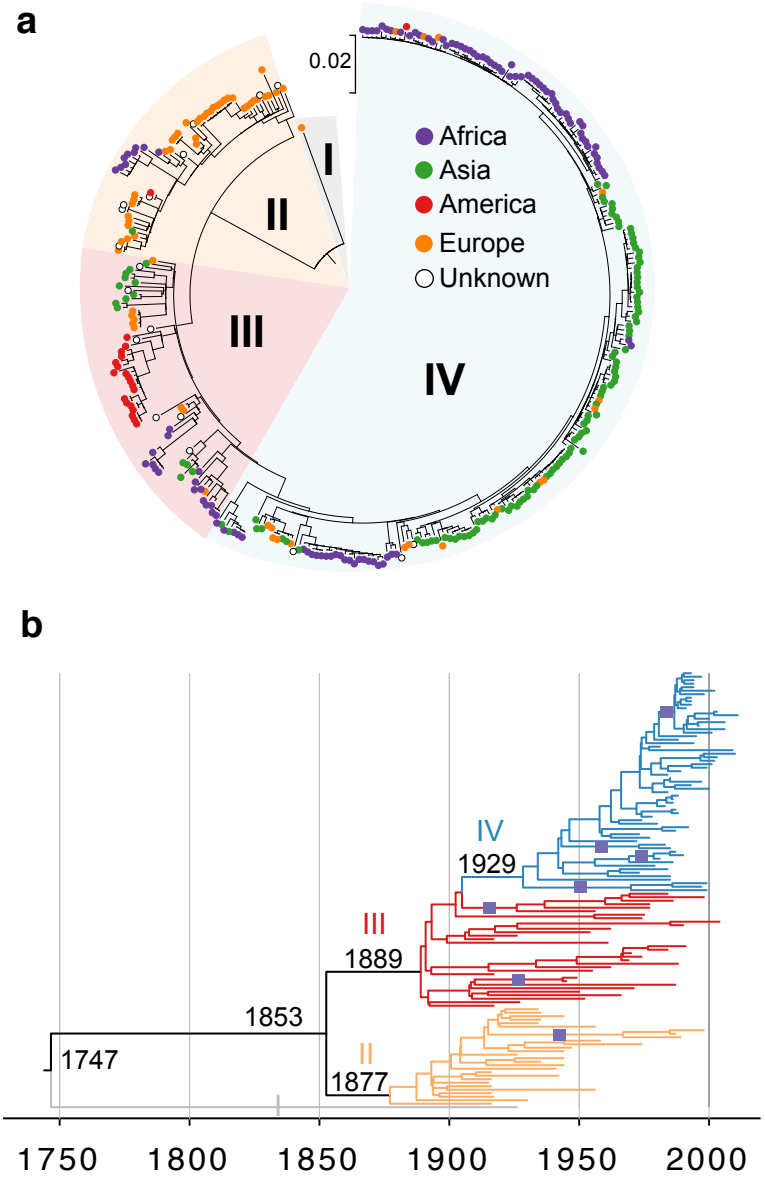

Figure 1.5: Population structure of *S. dysenteriae*. Figure adapted from Njamkepo *et al.*, 2016.[260] **a,** Maximum likelihood phylogeny of 235 *S. dysenteriae* genomes. Tips of the tree are coloured by continent, as per legend. Segments of tree are highlighted by lineage. **b,** Bayseian phylogeny of 125 *S. dysenteriae* genomes. Branches are coloured by lineage, with dates of emergence at each major node. Purple squares show intercontinental transmission events.

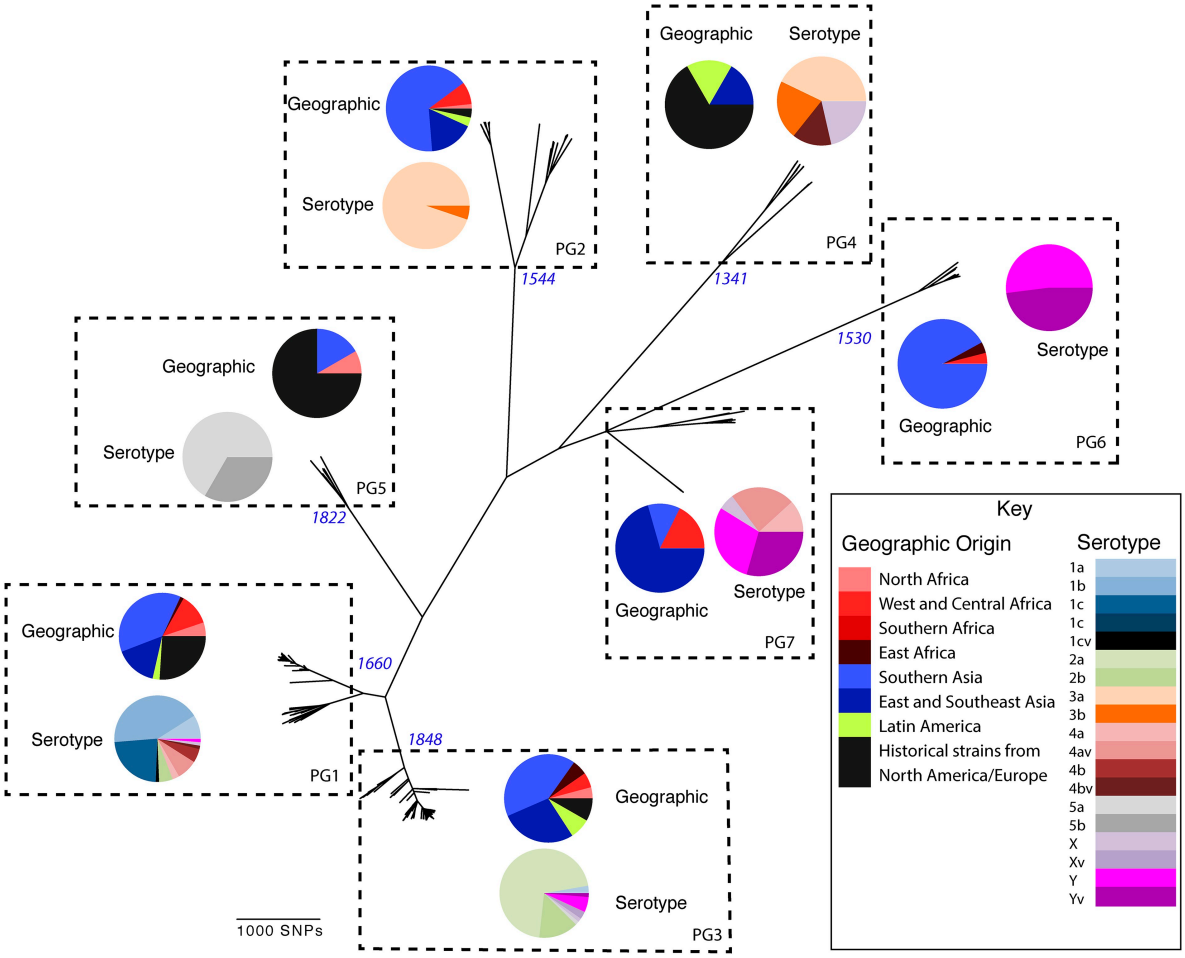

Figure 1.6: Maximum likelihood phylogeny for *S. flexneri* isolates including serotypes 1–5, X and Y produced from the results of mapping sequence reads against the genome of *S. flexneri* 2a strain 301, with recombination removed. Reproduced from Connor *et al.*, 2015.[264] Original legend: "Phylogenetic groups (PGs) determined by Bayesian analysis of population structure clustering are boxed within dotted lines, with the geographic and serotype composition of isolates in each PG being inlaid as pie charts."

1.5 Aims of this project

Since the advent of high throughput, short read sequencing, significant leaps have been made in our understanding of bacterial genome evolution. IS are frequently overlooked in these high throughput genomic studies that regularly analyse hundreds, or thousands, of bacterial genomes. This thesis aims to develop new software to enable detection of IS from short read data.

As IS are present in many copies throughout a genome, assembling across these regions is difficult. Several studies have examined IS in an experimental setting, investigating the specifics of IS within a few colonies of bacteria (section 1.2.2), but to date there are very few studies investigating the dynamics of IS within whole clones or populations of a species. A greater understanding of the dynamics of IS within different bacterial pathogen populations, and how they influence the evolution of that population, is still required.

Identifying IS from large bacterial datasets will aid understanding of the movement of mobile elements that carry important biological genes, such as antibiotic resistance genes. Additionally, IS have played a significant role in the evolution of *Shigella* from *E. coli*, however not much is known about the distribution and total burden of different IS species amongst the different *Shigella* species. *Shigella* have undergone several evolutionary bottlenecks, which have contributed to significant gene loss within their genomes. This thesis aims to investigate the IS species present in *Shigella*, their burden across evolutionary time, and their contribution to gene loss. Three different species of *Shigella*, *S. sonnei*, *S. dysenteriae* and *S. flexneri*, are examined.

i) The first aim of this project was to develop a new method for detecting IS within genomes from short read data. The resulting tool, ISMapper, was validated against several different datasets, including both simulated and real short read data.

ii) Secondly, this thesis aims to investigate how IS influence the spread of antibiotic resistance genes in three different bacterial species - *Salmonella* Kentucky, *S.* Typhi and *A. baumannii*.

iii) Finally, this thesis aims to investigate IS dynamics within a globally distributed pathogen, *Shigella*.

Chapter 2

Introducing ISMapper

2.1 Introduction

IS are important to the evolution and adaptation of bacterial pathogens. To date, the majority of bacterial genomic studies have primarily focused on mutations arising from SNPs. Despite the fact that IS can cause large-scale mutations and rearrangements in bacterial genomes, high throughput genomic studies have not traditionally studied them due to the difficulty of detecting them from short read data. There are few validated tools that are designed to detect IS from short read data. Most tools developed to detect transposable elements from genomic data were not specifically designed for use with bacterial data sets, and those developed to detect larger structural variation typically do not perform well at detecting IS. This chapter introduces a novel software tool, ISMapper, that is specifically designed for rapid, accurate detection of IS sites and their orientation in bacterial genomes using short read data.

As input, ISMapper requires paired-end short reads, an IS query, and either a reference genome or assembly. It can be used to detect IS positions relative to a reference (typing mode) or to assist with the resolution of complex regions in assemblies (improvement mode). ISMapper chains together existing tools (BWA[265], SAMtools[266] and BLAST[267]) in a Python framework to detect IS.

The paper below (Section 2.2) describes the validation and implementation of ISMapper, which was performed using both real and simulated data. In the validation with simulated data, reads were simulated from finished genomes and ISMapper was used to detect known IS within them. Validation with real data was performed in two parts. Firstly, finished genomes where IS positions were known and for which Illumina read were available were used to validate ISMapper. Secondly, Illumina data from several unpublished *A. baumannii* genomes were screened for ISAba1 and ISAba125 using ISMapper, and IS positions were confirmed by PCR by my collaborators, Mohammad Hamidian and Ruth Hall at the University of Sydney. Finally, ISMapper was used to screen for IS*6110* in 138 publicly available *M. tuberculosis* genomes to demonstrate its utility for analysing large genomic data sets. This analysis produced novel insights into the evolution of IS within an important bacterial pathogen.

The ISMapper paper included comparison with, and discussion of, related tools that were available at the time of publication. Since then, three additional tools have been released: ITIS, ISseeker and ISQuest. Discussion of these tools is provided in section 2.3, including results and benchmarking using the same data sets for ISMapper validation.

2.2 Publication

Hawkey *et al. BMC Genomics* (2015) 16:667
DOI 10.1186/s12864-015-1860-2

BMC Genomics

SOFTWARE Open Access

ISMapper: identifying transposase insertion sites in bacterial genomes from short read sequence data

Jane Hawkey[1,2*], Mohammad Hamidian[3], Ryan R. Wick[1], David J. Edwards[1], Helen Billman-Jacobe[2], Ruth M. Hall[3] and Kathryn E. Holt[1]

Abstract

Background: Insertion sequences (IS) are small transposable elements, commonly found in bacterial genomes. Identifying the location of IS in bacterial genomes can be useful for a variety of purposes including epidemiological tracking and predicting antibiotic resistance. However IS are commonly present in multiple copies in a single genome, which complicates genome assembly and the identification of IS insertion sites. Here we present ISMapper, a mapping-based tool for identification of the site and orientation of IS insertions in bacterial genomes, directly from paired-end short read data.

Results: ISMapper was validated using three types of short read data: (i) simulated reads from a variety of species, (ii) Illumina reads from 5 isolates for which finished genome sequences were available for comparison, and (iii) Illumina reads from 7 *Acinetobacter baumannii* isolates for which predicted IS locations were tested using PCR. A total of 20 genomes, including 13 species and 32 distinct IS, were used for validation. ISMapper correctly identified 97 % of known IS insertions in the analysis of simulated reads, and 98 % in real Illumina reads. Subsampling of real Illumina reads to lower depths indicated ISMapper was able to correctly detect insertions for average genome-wide read depths >20x, although read depths >50x were required to obtain confident calls that were highly-supported by evidence from reads. All ISAba1 insertions identified by ISMapper in the *A. baumannii* genomes were confirmed by PCR. In each *A. baumannii* genome, ISMapper successfully identified an IS insertion upstream of the *ampC* beta-lactamase that could explain phenotypic resistance to third-generation cephalosporins. The utility of ISMapper was further demonstrated by profiling genome-wide IS*6110* insertions in 138 publicly available *Mycobacterium tuberculosis* genomes, revealing lineage-specific insertions and multiple insertion hotspots.

Conclusions: ISMapper provides a rapid and robust method for identifying IS insertion sites directly from short read data, with a high degree of accuracy demonstrated across a wide range of bacteria.

Keywords: Insertion sequence (IS), Bacteria, Genomics, Short read analysis, Tuberculosis, Antimicrobial resistance

Background

An insertion sequence (IS) is a small transposable element that encodes the proteins required for its own transposition. The ISfinder database [1] currently contains over 500 distinct IS. During transposition some ISs create direct repeats, or target site duplications, in the sequences into which they are integrating. The presence and length of these duplications vary widely between ISs and are characteristic of individual IS [2]. Rates of transposition vary between ISs and host species, but are frequently in the order of the rate of nucleotide substitutions, making IS activity one of the more dynamic evolutionary forces at play in many bacterial genomes. The movement of ISs can also have functional consequences for bacterial genomes. ISs have been implicated in large changes to genome structure, by expanding in copy number in microbial

* Correspondence: hawkey.jane@gmail.com
[1]Department of Biochemistry and Molecular Biology, Bio21 Molecular Science and Biotechnology Institute, The University of Melbourne, Parkville, VIC 3010, Australia
[2]Faculty of Veterinary and Agricultural Science, The University of Melbourne, Parkville, VIC 3010, Australia
Full list of author information is available at the end of the article

genomes, with subsequent loss of ISs resulting in inactivation of genes, pseudogene formation, mediating deletion of intervening sequences between two copies of the IS, or rearrangements of the genome [3].

In addition, IS insertions upstream of protein coding sequences can result in their enhanced expression, leading to different phenotypes depending on the function of the over-expressed gene. There are several known examples of IS-mediated gene expression leading to clinically important increases in antimicrobial resistance. For example, increased resistance to fluoroquinolones such as ciprofloxacin can result from the insertion of IS*1* or IS*10* upstream of the *acrEF* efflux pump in *Salmonella* Typhimurium [4], or the insertion of IS*186* upstream of the *acrAB* efflux pump in *Escherichia coli* [5]. In *Acinetobacter baumannii*, insertion of ISAba1 or ISAba125 upstream of the intrinsic beta-lactamase *ampC* can cause resistance to third generation cephalosporins including ceftazidime and cefotaxime [6, 7]. Insertions of the same IS in nearby locations can generate a composite transposon, capable of mobilizing the intervening sequence and transferring it to new genomic locations. For example, the composite transposon Tn*6168* was generated spontaneously via insertions of ISAba1 on either side of *ampC*, including one copy of ISAba1 that upregulates *ampC* expression [8]. Tn*6168* has then transferred into different *A. baumannii* backgrounds, conferring horizontally-acquired resistance to third generation cephalosporins [8].

IS insertions also result in the upregulation of virulence genes in clinically important human pathogens. For example, an outbreak of tuberculosis in Spain in the 1990s was associated with the B strain of *Mycobacterium bovis* carrying an insertion of IS*6110* in the promoter region of the virulence gene *phoP*, resulting in its upregulation [9]. In *Neisseria meningitidis*, insertion of IS*1301* in the middle of the capsule locus has been shown to cause increased expression of operons on either side of the IS, contributing to protection from the human immune system and enhanced pathogenicity [10]. ISs have also been shown to enhance niche adaptation in bacteria, for example IS*1247* insertion upstream of *dhlB* in *Xanthobacter autotrophicus* results in increased resistance to bromoacetate [11]. This region has also been mobilised by the IS and transferred to a plasmid [11]. In *E. coli*, IS*3* has been shown to up-regulate threonine expression, allowing the bacteria to adapt to a low-carbon environment and utilise threonine as its sole carbon source [12].

The profiling of IS insertion patterns has been used for typing purposes in numerous bacterial species of importance to human health. For example, copy number and position of IS*200* in *Salmonella enterica* [13], IS*6110* in *Mycobacterium tuberculosis* [14], IS*1004* in *Vibrio cholerae* [15] and ISAba1 in *A. baumannii* [16] has been used to profile these bacterial pathogens, allowing the identification and tracking of distinct subtypes. To date, IS-based typing schemes for various bacteria have relied on digesting the genome followed by either hybridizing IS probes to fragments in a gel or PCR probing [13–15]. The detection of precise insertion sites can be achieved using PCR, and may be done for typing purposes [17] or for the detection of functionally important insertions [7, 9].

With the advent of cheap high-throughput short-read sequencing, whole genome sequencing (WGS) of bacteria is increasingly common and is replacing traditional methods for characterizing and typing bacterial genomes. Unfortunately the detection of ISs is complicated wherever read lengths are shorter than the length of the IS, as is the case for platforms that are currently most widely used – Illumina and Ion Torrent. IS insertion sites can readily be identified in finished bacterial genomes or in draft assemblies of genomes with single-copy ISs, using tools such as nucleotide BLAST or ISfinder [1]. However where multiple copies of the same IS are present within a single genome (including on the chromosome and/or plasmids), this complicates assembly of short-read data and makes IS insertion sites difficult to identify reliably. The IS detection problem can be resolved using long-read sequencing technologies such as the SMRT Cell (Pacific Biosciences) or MinION (Oxford Nanopore) platforms; however given the relative cost efficiency and reliability of short-read sequencing, together with the current widespread use of Illumina for bacterial WGS and wealth of available short-read data for clinically important bacteria, there remains a need for a simple tool to identify IS insertion sites from short-read data.

Several studies report the use of mapping-based approaches to identify IS insertion sites from bacterial short-read data [18, 19], however none provide software code or validation of the approach used. There are tools available for detecting transposons or structural variation in genomes, for example MindTheGap [20], BreakDancer [21], and Mobster [22], however these do not perform well in the identification of IS in bacterial genomes nor were they designed to do so. Some programs could potentially be used for this purpose, such as RelocaTE [23] and RetroSeq [24], however these require additional input or prior knowledge about the IS which may not always be available. TIF (Transposon Insertion Finder) [25] and *breseq* [26] could potentially be used for the detection of IS insertion sites in bacterial genomes, however they were not designed specifically for this purpose and did not perform well on our data sets (see Results).

Here we present a rapid and robust tool for accurate detection of ISs, including insertion site and orientation,

direct from short-read data. The method is freely available in the form of open-source code called ISMapper, and here we validate its use via analysis of simulated and real short-read data from a range of ISs and bacterial species. ISMapper requires short reads and query IS sequences as input, and can be used either for typing against a reference genome or to assist with manual resolution of complex short-read assemblies.

Implementation

An overview of the ISMapper workflow is shown in Fig. 1. ISMapper takes as input: (i) a set of paired end Illumina reads for an isolate of interest, (ii) an IS query sequence in fasta format, and (iii) either a reference genome (for typing) or an assembly of the read set (for assembly improvement), in GenBank or FASTA format (Fig. 1a). Paired end Illumina reads are mapped to the IS query sequence using BWA-MEM (v0.7.5a or later) [27]. From the resulting alignment file (SAM format), unmapped reads whose pairs map to the end of the IS query sequence (that is, reads representing the sequences directly flanking the IS) are extracted using SAMtools view (v0.1.19 or later) [28] to retrieve reads based on SAM flags (Fig. 1b). Specifically, left flanking reads (taking input sequence as left to right) are extracted using flag '–f 36' and right flanking reads are extracted using flag '–F 40 –f 4' and stored in separate BAM files, which are then converted to FASTQ format using BedTools (v2.20.1) [29]. In addition, Samblaster (v0.1.21) [30] is used to extract from the SAM file any

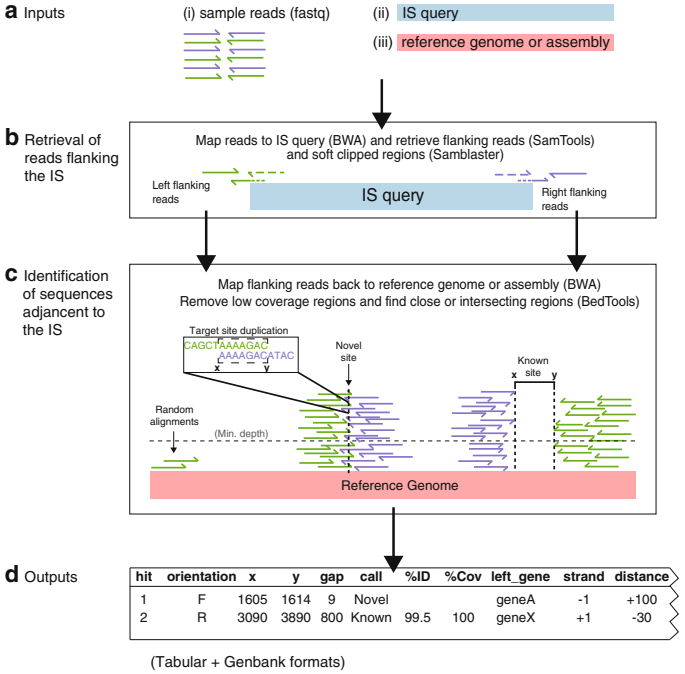

Fig. 1 Workflow for ISMapper. **a** Inputs are reads (fastq format) and an IS query (fasta), as well as either a reference genome to compare to or an assembly of the reads (fasta or genbank). **b** Reads are mapped to the IS query using BWA (*dashed lines*) and their pairs (i.e., *flanking reads, solid lines*) are retreived from the resulting SAM file using mapping flags (SAMTools). The unmapped component of soft clipped reads (*solid + dashed lines*), identified from the SAM file using Samblaster) are also retrieved using BioPython. **c** Flanking and softclipped read sequences are then mapped to either the reference genome or the read assembly (BWA), and the final mapping output is then filtered for depth to remove low coverage regions. Left and right end blocks are extracted from the resulting BAM file (using BedTools) and compared to one another to find either intersecting regions, indicating a novel IS position, or close regions, indicating a known IS position in the reference. In a novel position, overlapping sequences from the left and right ends most likely indicate the target site duplication generated during IS transposition (*zoomed in section*). **d** Results are tabulated, indicating the position and orientation of each site, whether it is novel or known, and information about the genes flanking the insertion site

reads that map to the end of the IS and extend into the neighbouring sequence (i.e., "soft clipped" reads, Fig. 1b). The resulting FASTQ file is filtered using BioPython to extract the soft clipped portion of reads, where those sequences fit a specified size range (default 5–30 bp). The resulting sequences are sorted into left and right flanking sequences; these are each mapped separately to the reference genome or assembly using BWA-MEM, to identify the location(s) of the query IS in the genome under analysis (Fig. 1c). Insertion site information is extracted from the resulting alignments using BedTools (coverage command) to summarise coverage of the reference by left and right flanking reads; these are filtered by read depth (default, minimum read depth ≥6x) to minimize false positive hits, and regions that overlap or are separated by a short distance (default, ≤100 bp) are merged using BedTools (merge command). Pairs of left and right flanking regions that likely represent either side of the same IS insertion are identified on the basis of positional information, using BedTools (intersect and closest commands). Left and right regions that overlap are considered to indicate a novel IS insertion not present in the reference, with the overlap resulting from target site duplication arising during IS transposition (Fig. 1c, novel site). Coordinate x refers to the left side of the target site duplication, and y refers to the coordinate of the right side of the target site duplication. In cases where the left and right flanking regions are extremely close but do not overlap, x refers to the inner end of the left-most region, and y refers to the inner end of the right-most region (as per a known site, see below). Where left and right regions are separated by a sequence that is approximately the length of the IS query, the intervening sequence is extracted and compared to the IS query using nucleotide BLAST+ (v2.2.25 or later) [31] to confirm whether this is a known insertion site that is present in the reference (Fig. 1c, known site). In this case, coordinate x refers to the inner end of the left-most block, and y refers to the inner end of the right-most block. Coordinate x is always the smallest number and y always the largest, regardless of IS orientation.

Extensive testing of ISMapper revealed that it was sometimes unable to resolve IS positions that were adjacent to a repeat region (segments of DNA that were repeated multiple times around the genome; see Results). This is because when the IS-flanking reads were mapped back to the reference genome, those that belonged to the neighbouring multi-copy sequence were randomly assigned by BWA-MEM to the various locations of the repeat sequence, resulting in low read depth at the 'true' IS-adjacent copy of the multi-copy sequence, which can fall below the minimum depth filter (Fig. 2). In such cases, the sequence on the other side of the IS is usually not a multi-copy sequence and thus does not suffer the same problem, and so is usually identified as a confident IS-flanking region without a corresponding partner region (Fig. 2, purple block). Therefore, when ISMapper identifies an un-partnered IS-flanking region, it checks the original alignments for evidence of a nearby low-coverage partner region that failed to pass the depth filter and returns this as a potential but uncertain IS location, indicated by a '?' character in the results table (see example of low-coverage partner region in Fig. 2, green block).

ISMapper generates two main output files summarizing the results: (i) a GenBank file of the reference sequence, annotated with the IS-flanking regions and (ii) a table indicating the locations and characteristics of each IS-flanking region identified (Fig. 1d). The table includes details of the location of the IS insertions (indicated by coordinates x and y); the distance between the left and right flanking regions (where a negative number indicates an overlap of left and right regions, indicating the size and sequence of the target site duplication); a call as

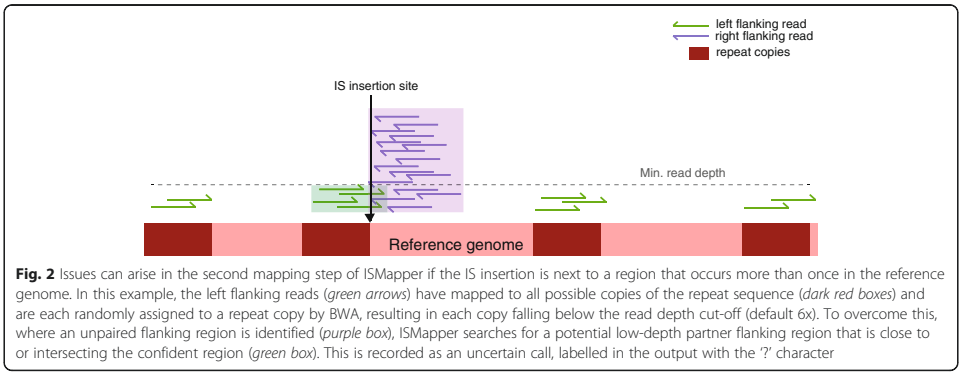

Fig. 2 Issues can arise in the second mapping step of ISMapper if the IS insertion is next to a region that occurs more than once in the reference genome. In this example, the left flanking reads (*green arrows*) have mapped to all possible copies of the repeat sequence (*dark red boxes*) and are each randomly assigned to a repeat copy by BWA, resulting in each copy falling below the read depth cut-off (default 6x). To overcome this, where an unpaired flanking region is identified (*purple box*), ISMapper searches for a potential low-depth partner flanking region that is close to or intersecting the confident region (*green box*). This is recorded as an uncertain call, labelled in the output with the '?' character

to whether the insertion is present in the reference or is a novel insertion site (and, where the insertion site is present in the reference, the percent coverage and sequence homology with the IS query); and details of the gene(s) closest to the IS insertion site (including locus tag, product, gene name and distance from the IS to the start codon). Insertions are also marked to indicate less confident calls. A '*' indicates an imprecise hit; i.e., where the gap between left and right regions is larger than expected for a novel insertion, but is not consistent with an IS insertion at that location in the reference. A '?' indicates an uncertain hit, where only one end (left or right of the predicted insertion) passes the minimum read depth threshold; this often occurs when the IS is inserted within or adjacent to a multi-copy sequence, as described above (Fig. 2). When run in assembly improvement mode, the table produced is simpler and indicates which contigs are predicted to end adjacent to the IS (indicating left or right orientation), assisting the user to decide whether some contigs could be joined together based on the available IS evidence.

ISMapper is lightweight code – a test run on a laptop computer (MacBook Air) with 8GB of RAM and a 1.3GHz i5 processor was able to analyse a read set comprising 2.5 million 100 bp paired-end reads in approximately ten minutes for a single IS query. Because ISMapper analyses each read set and query IS independently, screening of multiple read sets and query IS can be easily performed in parallel across multiple cores. To facilitate easy compilation of results from multiple jobs, ISMapper includes a Python script to cross-tabulate results from multiple read sets, generating a single summary table per query IS (script 'compiled_table.py').

Results and discussion

Validation of IS detection using simulated reads

Nine publicly available finished genomes from a variety of bacterial genera, and including both chromosomes and plasmids, were downloaded from NCBI (Table 1). ISfinder [1] was used to identify the ISs present in each finished genome sequence. All sequences that had >50 % identity to a sequence in ISfinder and were present in at least two copies were tested using ISMapper (with the query IS being sourced from curated references in ISfinder). Nucleotide BLAST+ was used to confirm the precise locations and orientations for each query IS in all genomes (total 251 insertions of 17 distinct IS, see Table 1). Short reads (100 bp) were simulated from each genome sequence using the wgsim command in SAMtools (v0.1.19), with default parameter settings.

ISMapper was run with default parameter settings on each combination of genome, query IS and simulated reads. ISMapper was able to accurately locate each IS position and its orientation (ranging between 2 and 61 positions per genome) for the majority of genomes (Table 1). In total, 97 % of IS insertions were correctly detected, with a false positive rate of 2.1 % ($n = 6$). The exceptions occurred in three genomes (*K. pneumoniae* plasmid pNDMAR, *Y. pestis* CO92 and *E. coli* O157:H7), in which ISMapper correctly identified 151 IS insertion sites and failed to identify nine (94 % detection). Closer inspection revealed that the missed IS were each located next to multi-copy repeat sequences, complicating the second mapping step as discussed above and outlined in Fig. 2. Switching on reporting of all alignments above a mapping score threshold of 30 (–a and -T 30 in BWA-MEM) enabled the detection of a further IS*100* site in *Y. pestis*. By default this option is turned off in ISMapper as it tends to create noise in the mapping, making it more difficult to distinguish true and false positives;

Table 1 Validation of ISMapper using simulated reads

Isolate	Accession	IS	Found	Orientation
S. Typhi CT18	NC_003198	IS*200*	26/26	26/26
		IS*1*	3/3	3/3
S. Typhimurium LT2	NC_003197	IS*200*	6/6	6/6
		ISSty2[b (1)]	2/2	2/2
		IS*1351*[b (2)]	2/2	2/2
S. Typhi plasmid pHCM1	NC_003384	IS*26*	4/4	4/4
		ISVsa5	2/2	2/2
		IS*1*	5/5	5/5
S. Paratyphi plasmid pAKU_1	AM412236	IS*26*	5/5	5/5
		ISVsa5	2/2	2/2
		IS*1*	7/7	7/7
K. pneumoniae plasmid pNDMAR	JN420336	IS*3000*	3/3	3/3
		IS*26*[a]	3/5	3/5
		ISEcp1	2/2	2/2
Yersinia pestis C092	NC_003143	IS*100*[a]	43/44	43/44
		IS*1661*[a, b (1)]	7/8	7/8
		IS*1541*[a]	61/64	61/64
Escherichia coli O104:H4	NC_018658	IS*1*	10/10	10/10
		IS*421*	4/4	4/4
		IS*609*	4/4	4/4
		ISEc1	4/4	4/4
		ISKpn26	4/4	4/4
E. coli O157:H7	NC_002695	IS*629*[a, b (1)]	17/18	17/18
		IS*609*	2/2	2/2
		ISEc1[b (1)]	4/4	4/4
		ISEc8[a]	9/10	9/10

[a]indicates ISMapper was unable to resolve some IS positions due to repeat regions
[b]indicates ISMapper incorrectly identified an insertion site, the number of these sites are indicated in brackets

however this can be useful if an IS site of interest is known or suspected to be flanked by further repeats.

Validation of IS detection using real Illumina read sets derived from isolates with finished genomes

Next we validated ISMapper using six finished genomes for which both Illumina read data and finished genomes were publicly available (Table 2). Each finished genome sequence was analysed with ISfinder [1] to identify query ISs for testing as described above, and nucleotide BLAST was used to confirm the precise locations and orientations of each IS in each genome. The resulting test set comprised 106 insertions of 14 query ISs. Using default settings, ISMapper was able to accurately identify each IS insertion site and its orientation, between 2 and 26 per genome, for the majority of genomes (Table 2). In total, 104 (98 %) IS insertions were correctly detected by ISMapper, with 3 IS insertions falsely detected (false positive rate of 2.5 %, $n = 3$). Three of four IS431mec insertions in *Staphylococcus aureus* TW20 were correctly detected, however the fourth was missed by ISMapper as it was flanked by another IS431mec and further repeat sequences. Two of three IS1 insertions in *Salmonella* Typhi CT18 were correctly detected however a third, located between *tviE* and *tviD*, was problematic. ISMapper identified the region flanking the IS at *tviE*, but did not detect any corresponding region in *tviD*. To investigate why, we mapped the entire Illumina read set to the CT18 chromosome reference sequence, which showed that there were no reads derived from the region between *tviD* to *tviA*. Therefore this region appears to have been deleted during culture in the laboratory prior to the extraction of DNA for Illumina sequencing. This region encodes the biosynthesis of the Vi capsule of *S.* Typhi, and is known to be lost sporadically during culture [32]. This illustrates that situations where one end of the IS is detected but the other is not can often be 'accurate' in the sense that the result reflects underlying structural variation in the genome, including potentially IS-mediated deletions.

Detection of antibiotic resistance-mediating IS insertions in *Acinetobacter baumannii*, confirmed by PCR

The genomes of seven ceftazidime resistant *A. baumannii* isolates, belonging to global clone (GC) 1, were sequenced via Illumina HiSeq to generate 100 bp paired end reads. Resistance gene screening of the Illumina data using SRST2 [33] and the ARG-Annot database [34] confirmed earlier PCR data indicating that none of these isolates carried acquired extended spectrum beta-lactamase (ESBL) genes that can confer resistance to third-generation cephalosporins. However, it is known that the insertion of ISAba1 upstream of the intrinsic chromosomally encoded *ampC* beta-lactamase gene can cause increased resistance to third-generation cephalosporins in *A. baumannii* [6].

We used ISMapper to screen for the ISAba1 query sequence (accession AY758396), sourced from ISfinder [1]. Using default parameters, ISMapper identified ISAba1 insertions in all seven GC1 genomes. IS positions were assessed relative to the finished genome sequence of *A. baumannii* GC1 reference A1 (accession CP010781). ISMapper found between 3 and 5 ISAba1 insertions in each GC1 isolate, including an insertion upstream of

Table 2 Validation of ISMapper using real Illumina reads for which finished genomes were also available

Isolate	Genome accession	FastQ accession	IS	Found	Orientation
Streptococcus suis P1/7	AM946016	ERR225612	ISSu3	4/4	4/4
			ISSu4[b (2)]	2/2	2/2
Staphylococcus aureus TW20	NC_017331	ERR043367	ISSep3	3/3	3/3
			IS*256*	8/8	8/8
			IS431mec[a]	3/4	3/4
			IS*1181*	2/2	2/2
Klebsiella pneumoniae NJST258_1	CP006923	SRR1166975	ISKpn1	5/5	5/5
			IS5B	8/8	8/8
			IS903B[b (1)]	2/2	2/2
			IS*1294*	3/3	3/3
			ISKpn18	2/2	2/2
			ISKpn26	7/7	7/7
S. Typhi CT18	NC_003198	ERR343331	IS*200*	26/26	26/26
			IS*1*[a]	2/3	2/3
S. Typhi Ty2	AE014613	ERR343332	IS*200*	26/26	26/26

[a]indicates ISMapper was unable to resolve some positions due to repeat regions
[b]indicates ISMapper incorrectly identified an insertion site, the number of these sites are indicated in brackets

ampC in all 7 genomes that was in the orientation required to induce upregulation and explain the observed cephalosporin resistance phenotype (Fig. 3). In addition, out of 29 total ISAba1 insertions, ISMapper was able to correctly identify 26 target site duplications (9 bp in the case of ISAba1). All ISAba1 insertions were novel compared to the reference genome A1 (Fig. 3) and were confirmed using PCR, as described in [6].

Impact of read depth on ISMapper performance

To test the effect of read depth on the performance of ISMapper, each of the seven GC1 *A. baumannii* read sets were randomly subsampled to depths of approximately 10x, 15x, 20x, 25x, 50x, 75x and 100x, with ten replicates per depth level per read set. ISMapper was then run using default settings to screen for ISAba1 insertions. The results indicated that at mean genome-wide read depths of approximately 20x, ISMapper was able to identify 95 % of insertions correctly (Fig. 4). However, all of these calls were either imprecise (gap size larger than expected) or uncertain (high coverage end paired with a low coverage end). An average genome-wide read depth of ~50x was required to find all insertions, with confident calls for >60 %, however there was clearly some variation depending on read quality (Fig. 4). To achieve 100 % detection with high confidence, average genome-wide read depths of >75x were required (Fig. 4).

Comparison of ISMapper with TIF and *breseq*

The seven *A. baumannii* GC1 genomes were used to test both *breseq* [26] and TIF (Transposon Insertion Finder) [25]. *breseq* uses split read mapping to a reference genome along with statistical models to determine new junctions and deletions in the isolates of interest. As input, *breseq* takes paired end reads in FASTQ format, and a reference genome in Genbank format. The *breseq* manual indicates that new insertions of mobile elements can determined by looking for 'JC JC' evidence types in the final html output. All seven *A. baumannii* isolates were screened using default parameters and the reference genome A1 (accession CP010781). In all cases, *breseq* was unable to identify any mobile element insertions, including no structural variation at the known ISAba1 insertion sites, although many other types of structural variation were detected.

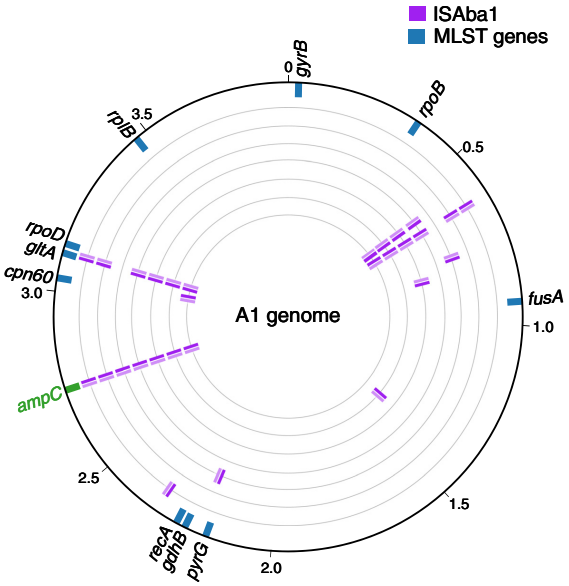

Fig. 3 ISAba1 positions detected in seven *Acinetobacter baumannii* genomes. The A1 reference genome is indicated by the outer black circle, tick marks indicate genome coordinates (in Mbp), genes targeted in the two available MLST schemes are marked in blue for orientation purposes. ISAba1 detected within the individual test genomes are shown on inner rings (*purple*), orientation of the IS is indicated by shading (*dark, left end; light, right end*). Novel IS insertions were identified upstream of the beta-lactamase *ampC* (*green*) in all isolates, explaining the observed phenotypic resistance to third generation cephalosporins

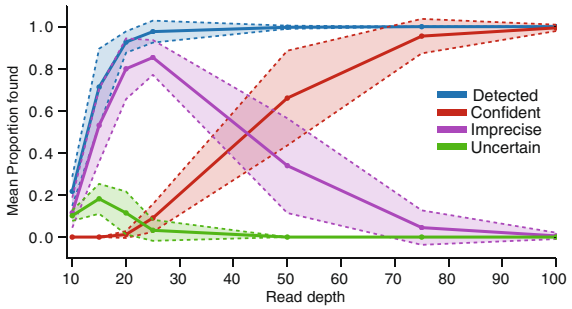

Fig. 4 ISAba1 detection rate as a function of read depth, for seven *A. baumannii* GC1 genomes sequenced with Illumina. Mean (*lines*) and standard deviation (*shaded areas*) proportions of IS insertions correctly detected per genome, amongst 70 replicate read sets sampled at each depth level (7 read sets x 10 replicates each at read depths 10x, 15x, 20x, 25x, 50x, 75x, 100x). Blue, IS insertion site detected; red, detected with high confidence; purple, detected with low precision (larger than expected gap size between left and right flanking regions, indicated with '*' in output); green, detected with low confidence (high-depth evidence for one side only, low-depth evidence for the other, indicated with '?' in output)

TIF requires as input paired end reads (FASTQ format), the head and tail sequences (approximately 17 bp) of the IS of interest as well as the size of the target site duplication the IS makes during transposition. TIF uses regular expressions to search for the head and tail sequences in the reads, and these reads are then extracted and grouped by their target site duplications. Unfortunately, following communication with the authors, we were unable to get TIF to output any results using our data. Other disadvantages of TIF are the requirements to (i) specify the size of target site duplications (which not all IS make and is not always known), (ii) manually extract subsequences of the IS rather than inputting the complete sequence, and (iii) manually edit a Perl script in order to specify inputs to the program.

Example use case: exploration of IS*6110* insertions in *Mycobacterium tuberculosis*

While IS insertions are thought to be important for shaping the evolution of bacteria in a variety of ways, high-resolution comparative genomic studies of bacterial pathogens have largely ignored ISs due to the difficulties associated with accurate detection of insertion sites from high-throughput short read data. An important example is IS*6110* in *M. tuberculosis* [35]. Profiling of IS*6110* insertions using PCR and restriction fragment based polymorphism (RFLP) based methods has been reported for typing purposes [36], and specific insertions have been linked to clinically relevant changes in function including in outbreak strains [9, 37, 38] However while numerous studies have reported the genomic analysis of hundreds of *M. tuberculosis* isolates sequenced on the Illumina platform, these have not included analysis of IS*6110* insertions. Thus, to demonstrate the utility of ISMapper for comparative profiling of ISs in an important bacterial pathogen, we analysed the distribution of IS*6110* within 138 publicly available genomes representing the major lineages of *M. tuberculosis* [39]. Paired-end Illumina reads were downloaded from NCBI (ERP001731). A core genome phylogeny was generated from these reads by SNP (single nucleotide polymorphism) calling against reference genome H37Rv (accession NC_000962) (methods as described in [40]), followed by maximum likelihood phylogenetic inference on the SNP alignment using RAxML (GTR + G substitution model, 1000 bootstraps) to build a genome-wide phylogenetic tree. ISMapper was run with default settings to screen for insertions of IS*6110* (accession X17348) in each read set, relative to reference genome H37Rv.

A total of 392 unique IS*6110* insertion sites were identified by ISMapper, approximately one per 10 kbp of the 4.4 Mbp reference genome. The frequency of each insertion within each of the six main global lineages is shown in Fig. 5b. The data indicate multiple lineage-specific IS*6110* insertions in lineages 2–6, but none that were shared by multiple lineages, suggesting that IS*6110* insertions began to accumulate only after *M. tuberculosis* diverged into these distinct lineages. Isolates in the "modern" lineages 2–4 and in the West African lineage 5 had more IS*6110* insertions overall, with far fewer insertions observed in the "ancient" East and West African lineages 1 and 6 (Fig. 5c). Lineage 2, which includes the highly successful Beijing sublineage, had the highest number of IS*6110* although it was not the most common lineage in the collection ($n = 23$); it could be that these insertions contribute to the adaptive fitness of the Beijing lineage.

The spatial distribution of unique IS*6110* insertions within the *M. tuberculosis* genome (Fig. 5d) revealed

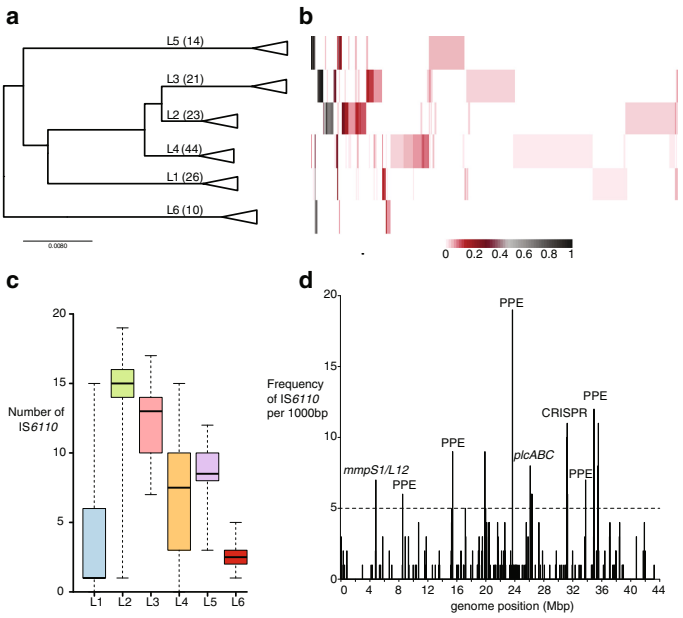

Fig. 5 Analysis of IS*6110* insertions in a diverse set of *M. tuberculosis* genomes. Analyses are based on ISMapper analysis of publicly available Illumina paired end data from 138 genomes. **a** Phylogenetic tree for *M. tuberculosis* based on genome-wide SNP calls, with midpoint rooting and collapsed to lineage level. Number of isolates analysed in each lineage (L) is indicated in brackets. **b** Heat map indicating the frequency of each IS*6110* insertion site (*columns*) detected in each lineage (*rows*), according to inset legend (i.e., a black cell indicates that the IS insertion site was detected in all isolates of the given lineage, white cell indicates that the insertion site was not detected at all in that lineage). **c** Boxplots show number of IS*6110* insertions detected per genome, for each lineage. Black line, median; boxes, interquartile range; whiskers, minimum and maximum values. **d** Histogram of IS*6110* insertions in 1000 bp windows along the *M. tuberculosis* chromosome. Dashed line, threshold for defining insertion hotspots

several clusters of insertions detected by ISMapper. Many of these clusters comprised multiple independent insertions into PE/PPE genes (which are surface-associated and interact with the host immune system), as well as the membrane associated proteins *mmpS1* and *mmpL12*. There was substantial clustering of IS*6110* insertions interrupting genes encoding the CRISPR machinery, which is involved in immunity to bacteriophage and other foreign DNA. Further, all three phospholipase genes, which are involved in virulence by inducing cell death in macrophages [41] and are encoded by the *plcABC* operon, contained multiple IS*6110* insertions detected by ISMapper. This locus is a known hotspot for IS*6110* insertions and has been shown to mediate deletions of segments of this region [42]. IS*6110* insertions upstream of *phoP*, which have been associated with upregulation and enhanced virulence in *M. tuberculosis* [9], were identified in multiple lineages (1 insertion in 6 lineage 2 genomes; singular insertions in one genome each in lineage 3 and 5) and may be indicative of positive selection for enhanced *phoP* expression and virulence. These findings from ISMapper analysis are consistent with those reported from PCR-based screens of smaller sets of isolates, but provide a more comprehensive picture of IS dynamics in *M. tuberculosis* that could be extended to much larger genomic data sets and other important pathogens.

Conclusions

ISMapper is a lightweight and reliable tool for the detection of IS insertion sites in bacterial genomes using high-throughput short-read sequencing data, which is now ubiquitous in microbial research and clinical investigations. ISMapper performed well on real and simulated data from 32 different ISs and 13 bacterial species, detecting all but the most complex instances involving multiple neighbouring IS insertions or other repeated sequences. ISMapper was able to detect antimicrobial

resistance-associated ISAba1 insertions in *A. baumannii*, with all sites detected by the program being subsequently confirmed by PCR. Compared to other tools such as *breseq* and TIF, ISMapper is ideal for detecting new positions for known ISs in bacterial genomes. In addition, ISMapper was able to rapidly produce a wealth of data on IS*6110* insertions in *M. tuberculosis*, allowing quick identification of lineage-specific insertions and specific regions enriched for insertions that may be functionally significant.

Availability and requirements

- **Project name:** ISMapper
- **Project home page:** https://github.com/jhawkey/IS_mapper
- **Programming language:** Python v2.7.5
- **Operating system(s):** platform independent, requires Python 2.7 and dependencies
- **Other requirements:** BioPython v1.63, BWA v0.7.12, SAMtools v1.1, Bedtools v2.20.1, BLAST+ v2.2.28, Samblaster v0.1.21
- **License:** Modified BSD

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
JH developed the code, analysed data and wrote the paper. KEH conceived the study and helped to draft the manuscript. HBJ participated in design and coordination of the study and contributed to data interpretation. RRW and DJE developed code. MH performed PCR and sequence analysis. RMH provided sequence data and isolates for validation and contributed to data interpretation. All authors read and approved the final manuscript.

Acknowledgements
This work was supported by the National Health and Medical Research Council of Australia (Fellowship #1061409 to KEH; Project Grant #1043830 to KEH and RMH) and the Victorian Life Sciences Computation Initiative (VLSCI, #VR0082).

Author details
[1]Department of Biochemistry and Molecular Biology, Bio21 Molecular Science and Biotechnology Institute, The University of Melbourne, Parkville, VIC 3010, Australia. [2]Faculty of Veterinary and Agricultural Science, The University of Melbourne, Parkville, VIC 3010, Australia. [3]School of Molecular Bioscience, The University of Sydney, Sydney 2006, Australia.

Received: 9 March 2015 Accepted: 18 August 2015
Published online: 03 September 2015

References
1. Siguier P, Perochon J, Lestrade L, Mahillon J, Chandler M. ISfinder: the reference centre for bacterial insertion sequences. Nucleic Acids Res. 2006;34(Database issue):D32–6.
2. Mahillon J, Chandler M. Insertion sequences. Microbiol Mol Biol Rev. 1998;62:725–74.
3. Siguier P, Gourbeyre E, Chandler M. Bacterial insertion sequences: their genomic impact and diversity. FEMS Microbiol Rev. 2014;38:865–91.
4. Olliver A, Vallé M, Chaslus-Dancla E, Cloeckaert A. Overexpression of the multidrug efflux operon *acrEF* by insertional activation with IS*1* or IS*10* elements in *Salmonella enterica* serovar *typhimurium* DT204 *acrB* mutants selected with fluoroquinolones. Antimicrob Agents Chemother. 2005;49:289–301.
5. Jellen-Ritter AS, Kern WV. Enhanced expression of the multidrug efflux pumps AcrAB and AcrEF associated with insertion element transposition in *Escherichia coli* mutants selected with a fluoroquinolone. Antimicrob Agents Chemother. 2001;45:1467–72.
6. Hamidian M, Hall RM. ISAba1 targets a specific position upstream of the intrinsic *ampC* gene of *Acinetobacter baumannii* leading to cephalosporin resistance. J Antimicrob Chemother. 2013;68:2682–3.
7. Hamidian M, Hancock DP, Hall RM. Horizontal transfer of an ISAba125-activated *ampC* gene between *Acinetobacter baumannii* strains leading to cephalosporin resistance. J Antimicrob Chemother. 2013;68:244–5.
8. Hamidian M, Hall RM. Tn*6168*, a transposon carrying an ISAba1-activated *ampC* gene and conferring cephalosporin resistance in *Acinetobacter baumannii*. J Antimicrob Chemother. 2014;69:77–80.
9. Soto CY, Menéndez MC, Pérez E, Samper S, Gómez AB, García MJ, et al. IS*6110* mediates increased transcription of the *phoP* virulence gene in a multidrug-resistant clinical isolate responsible for tuberculosis outbreaks. J Clin Microbiol. 2004;42:212–9.
10. Uria MJ, Zhang Q, Li Y, Chan A, Exley RM, Gollan B, et al. A generic mechanism in *Neisseria meningitidis* for enhanced resistance against bactericidal antibodies. J Exp Med. 2008;205:1423–34.
11. Van Der Ploeg J, Willemsen M, Van Hall G, Janssen DB. Adaptation of *Xanthobacter autotrophicus* GJ10 to bromoacetate due to activation and mobilization of the haloacetate dehalogenase gene by insertion element IS*1247*. J Bacteriol. 1995;177:1348–56.
12. Aronson BD, Levinthal M, Somerville RL. Activation of a cryptic pathway for threonine metabolism via specific IS*3*-mediated alteration of promoter structure in *Escherichia coli*. J Bacteriol. 1989;171:5503–11.
13. Soria G, Barbé J, Gibert I. Molecular fingerprinting of *Salmonella typhimurium* by IS*200*-typing as a tool for epidemiological and evolutionary studies. Microbiologia. 1994;10:57–68.
14. Das S, Paramasivan CN, Lowrie DB, Prabhakar R, Narayanan PR. IS*6110* restriction fragment length polymorphism typing of clinical isolates of *Mycobacterium tuberculosis* from patients with pulmonary tuberculosis in Madras, South India. Tuber Lung Dis. 1995;76:550–4.
15. Bik EM, Gouw RD, Mooi FR. DNA fingerprinting of *Vibrio cholerae* strains with a novel insertion sequence element: a tool to identify epidemic strains. J Clin Microbiol. 1996;34:1453–61.
16. Adams MD, Chan ER, Molyneaux ND, Bonomo RA. Genomewide analysis of divergence of antibiotic resistance determinants in closely related isolates of *Acinetobacter baumannii*. Antimicrob Agents Chemother. 2010;54:3569–77.
17. Suzuki M, Matsumoto M, Hata M, Takahashi M, Sakae K. Development of a rapid PCR method using the insertion sequence IS*1203* for genotyping Shiga toxin-producing *Escherichia coli* O157. J Clin Microbiol. 2004;42:5462–6.
18. Doig KD, Holt KE, Fyfe JA, Lavender CJ, Eddyani M, Portaels F, et al. On the origin of *Mycobacterium ulcerans*, the causative agent of Buruli ulcer. BMC Genomics. 2012;13:258.
19. Doughty EL, Sergeant MJ, Adetifa I, Antonio M, Pallen MJ. Culture-independent detection and characterisation of *Mycobacterium tuberculosis* and *M. africanum* in sputum samples using shotgun metagenomics on a benchtop sequencer. PeerJ. 2014;2:e585.
20. Rizk G, Gouin A, Chikhi R, Lemaitre C. MindTheGap: integrated detection and assembly of short and long insertions. Bioinformatics. 2014;30:3451–7.
21. Chen K, Wallis JW, McLellan MD, Larson DE, Kalicki JM, Pohl CS, et al. BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. Nat Methods. 2009;6:677–81.
22. Thung DT, de Ligt J, Vissers LE, Steehouwer M, Kroon M, de Vries P, et al. Mobster: accurate detection of mobile element insertions in next generation sequencing data. Genome Biol. 2014;15:488.
23. Robb SMC, Lu L, Valencia E, Burnette JM, Okumoto Y, Wessler SR, et al. The use of RelocaTE and unassembled short reads to produce high-resolution snapshots of transposable element generated diversity in rice. G3. 2013;3:949–57.
24. Keane TM, Wong K, Adams DJ. RetroSeq: transposable element discovery from next-generation sequencing data. Bioinformatics. 2013;29:389–90.
25. Nakagome M, Solovieva E, Takahashi A, Yasue H, Hirochika H, Miyao A. Transposon Insertion Finder (TIF): a novel program for detection of de novo transpositions of transposable elements. BMC Bioinformatics. 2014;15:71.
26. Barrick JE, Colburn G, Deatherage DE, Traverse CC, Strand MD, Borges JJ, et al. Identifying structural variation in haploid microbial genomes from short-read resequencing data using breseq. BMC Genomics. 2014;15:1039.

27. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25:1754–60.
28. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/Map format and SAMtools. Bioinformatics. 2009;25:2078–9.
29. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26:841–2.
30. Faust GG, Hall IM. SAMBLASTER: fast duplicate marking and structural variant read extraction. Bioinformatics. 2014;30:2503–5.
31. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. BMC Bioinformatics. 2009;10:421.
32. Holt KE, Parkhill J, Mazzoni CJ, Roumagnac P, Weill F-X, Goodhead I, et al. High-throughput sequencing provides insights into genome variation and evolution in *Salmonella* Typhi. Nat Genet. 2008;40:987–93.
33. Inouye M, Dashnow H, Raven L-A, Schultz MB, Pope BJ, Tomita T, et al. SRST2: rapid genomic surveillance for public health and hospital microbiology labs. Genome Med. 2014;6:90.
34. Gupta SK, Padmanabhan BR, Diene SM, Lopez-Rojas R, Kempf M, Landraud L, Rolain J-M: ARG-ANNOT, a new bioinformatic tool to discover antibiotic resistance genes in bacterial genomes. Antimicrob Agents Chemother 2014;58:212–20
35. McEvoy CRE, Falmer AA, van Pittius NCG, Victor TC, van Helden PD, Warren RM. The role of IS*6110* in the evolution of *Mycobacterium tuberculosis*. Tuberculosis. 2007;87:393–404.
36. van Embden JD, Cave MD, Crawford JT, Dale JW, Eisenach KD, Gicquel B, et al. Strain identification of *Mycobacterium tuberculosis* by DNA fingerprinting: recommendations for a standardized methodology. J Clin Microbiol. 1993;31:406–9.
37. Beggs ML, Eisenach KD, Cave MD. Mapping of IS*6110* insertion sites in two epidemic strains of *Mycobacterium tuberculosis*. J Clin Microbiol. 2000;38:2923–8.
38. Alonso H, Aguilo JI, Samper S, Caminero JA, Campos-Herrero MI, Gicquel B, et al. Deciphering the role of IS*6110* in a highly transmissible *Mycobacterium tuberculosis* Beijing strain, GC1237. Tuberculosis. 2011;91:117–26.
39. Comas I, Coscolla M, Luo T, Borrell S, Holt KE, Kato-Maeda M, et al. Out-of-Africa migration and Neolithic coexpansion of *Mycobacterium tuberculosis* with modern humans. Nat Genet. 2013;45:1176–82.
40. Holt KE, Thieu Nga TV, Thanh DP, Vinh H, Kim DW, Vu Tra MP, et al. Tracking the establishment of local endemic populations of an emergent enteric pathogen. Proc Natl Acad Sci U S A. 2013;110:17522–7.
41. Assis PA, Espíndola MS, Paula-Silva FW, Rios WM, Pereira PA, Leão SC, et al. *Mycobacterium tuberculosis* expressing phospholipase C subverts PGE_2 synthesis and induces necrosis and alevolar macrophages. BMC Microbiol. 2014;14:128.
42. Vera-Cabrera L, Hernández-Vera MA, Welsh O, Johnson WM, Castro-Garza J. Phospholipase region of *Mycobacterium tuberculosis* is a preferential locus for IS*6110* transposition. J Clin Microbiol. 2001;39:3499–504.

2.3 Tools published since ISMapper's publication

In addition to the two tools discussed in section 2.2, three new tools for the detection of transposase sites have been written since the publication of ISMapper. These tools are designed for either the detection IS specifically in prokaryotes, or transposases more generally in eukaryotes. This section details each of these tools and how they compare to ISMapper.

2.3.1 ITIS

ITIS is very similar to ISMapper, requiring paired end short reads, an IS query, and a reference genome[268]. ITIS was developed for detecting transposons in eukaryotes, and is able to detect transposase sites that are both homozygous and heterozygous. It uses a similar method to ISMapper, mapping reads to a transposase sequence and extracting the reads that either flank the transposase, or are clipped when mapped to the transposase. These reads are mapped to the reference genome to identify the transposase insertion sites, in a similar fashion to ISMapper. Overall, ITIS performed well on the validation datasets used to test ISMapper, finding all copies of ISAba1 in the test *A. baumannii* genomes. ITIS was computationally fast, with similar runtimes to ISMapper. However, the output files are problematic. ITIS only outputs the final BED files that describe the start and end of each transposase hit. This information is not tabulated into a simple format that is easy for the user to read or use in downstream analysis. Additionally, ITIS does not output contextual information for each hit, such as whether the hit is within a gene or in an intergenic region.

2.3.2 ISseeker

ISseeker uses a similar concept to ISMapper and ITIS, but instead of using a mapping approach, it uses an assembly-based approach[269]. ISseeker requires an IS query sequence and a reference genome to compare to, but instead of short reads, it uses a genome assembly. ISseeker uses BLAST to query the assembly for the IS of interest, detecting which contigs contain a full copy of the IS including flanking sequence, and which contigs contain only a partial match to the IS at the end of contigs. For the partial end matches, ISseeker extracts a flanking region next to the IS, and uses BLAST to compare this against the reference genome.

It then combines this information to compile a table showing the locations of each IS site, including whether the IS location is within a gene or in an intergenic region. As ISseeker uses an assembly-based approach, it is able to detect more complicated rearrangement events. However, ISseeker only outputs the final results in the SQL database format, rather than plain text, making downstream analysis more difficult. It also requires assemblies of genomes, making the initial step of preparing the data for analysis computationally intensive.

ISseeker was compared to ISMapper using the same seven *A. baumannii* genomes presented in the validation section of the paper, querying for ISAba1. In all cases ISseeker was able to correctly identify all ISAba1 insertion sites within each genome. However, it failed to detect the insertion of ISAba1 upstream of *ampC*, incorrectly identifying the insertion as within the *ampC* gene. As ISseeker relies on assemblies, identifying exact IS locations will be reliant on the quality of the assembly.

2.3.3 ISQuest

ISQuest takes a very different approach to the search of IS in bacterial genomes. Firstly, ISQuest does not require an IS query sequence[270]. ISQuest takes either short reads or assembled contigs as input, and uses these to query against its transposaseDB, which is generated by extracting all entries labelled as 'insertion sequence' or 'transposase' from a set of GenBank entries ISQuest downloads from NCBI. ISQuest then uses BLAST to detect which of these IS are present in the sample. As output, it produces two tables - one detailing the number of copies of each IS type found, and another giving the details of each IS hit, including whether it is a partial or complete IS.

In testing, ISQuest was able to identify 57 IS hits in one of the *A. baumannii* genomes used for validation with ISMapper. ISQuest was unable to identify the names of these IS, so the resulting fasta file of hits was given to the ISFinder database for identification. Many of the ISQuest hits were very small, and contained either no matches or very low confident matches in the curated ISFinder database. Out of the 57 hits found by ISQuest, eight of them were confidently identified by ISFinder (BLAST E value $< 10^{-5}$). Of these eight IS, one was ISAba1. However, ISQuest was unable to accurately identify the number of copies or locations of ISAba1 within the genome. Two contigs were identified as containing a copy of ISAba1. One of these was a partial hit, and the other was a single contig that contained only ISAba1 sequence.

2.4 Conclusion

In this chapter, a new tool was introduced for the detection of IS in short read data. It has been shown to be able to quickly detect IS in a wide variety of bacterial genomes with good specificity and sensitivity, using both simulated and real Illumina reads. Its usefulness for investigating how IS have contributed to the evolution of an important bacterial pathogen was demonstrated using publicly available *M. tuberculosis* data. Since the publication of ISMapper, other tools have been published that also aim to detect transposase sequences using next generation sequencing data, illustrating the interest in understanding how transposases impact the evolution of different organisms.

Five competing tools for detecting transposases in genome data have been published, including TIF and breseq (published prior to ISMapper and discussed in section 2.2) and ITIS, ISseeker and ISQuest (published after ISMapper and discussed in section 2.3). ITIS is very similar in concept and approach to ISMapper; however its lack of user friendly outputs make it difficult to use in high throughput genomic studies. ISseeker generally performed well and is useful for genomes that are only available as assemblies; however its reliance on assemblies makes it more computationally intensive when the goal is to analyse Illumina sequence data. It also appears to be less accurate than ISMapper at identifying precise insertion sites and their functional impact on genes. Finally, ISQuest took a very different approach, generating its own transposase database and using this database to detect all transposases in the genome. However, using our test data, it was unable to correctly identify the copy numbers, locations, and identities of the transposase genes it detected. Using our test set of seven *A. baumannii* genomes with experimentally validated ISAba1 positions, ISMapper was the only tool able to precisely detect all locations, gene contexts, and output user friendly tables, directly from short reads.

Chapter 3

Applications of ISMapper to study antimicrobial resistance

3.1 Introduction

Currently, known antibiotic resistance genes and resistance-associated SNPs are commonly detected using WGS. While phenotyping for AMR is currently the gold standard, increasingly WGS data is being used to predict antibiotic resistance phenotypes in clinical isolates, with the aim of informing treatment[271]. Commonly, sequencing data is applied as a surveillance method[272]. Short reads or assemblies can be screened *in silico* against databases containing antibiotic resistance genes, and used to infer the evolution of antibiotic resistance between and across pathogens[273]–[275]. Frequently, IS-mediated AMR is overlooked due to the difficulty of detecting IS from short read data. ISMapper (introduced in Chapter 2) is able to solve some of these technical issues, allowing the detection of IS-mediated resistance from short read data, which may be applied in the analysis of large datasets of bacterial genomes.

In this chapter, I will demonstrate how I have applied ISMapper to investigate various IS-mediated resistance mechanisms in four studies across two different bacterial species - *S. enterica* and *A. baumannii*. The ISMapper analyses presented here each formed part of larger studies with collaborators, and where published my contribution is acknowledged as a co-author. Here, I focus on the IS-related analyses that I conducted and the resulting insights into AMR. Firstly, ISMapper was used to reconstruct a complex genomic island encoding MDR in *Salmonella* Kentucky. In genomic surveillance studies of *S.* Typhi[276] and *A. baumannii*[69], ISMapper was used to investigate the location of IS-mediated AMR transposons; IS-mediated upregulation of the intrinsic beta-lactamase, *ampC*, was also explored in the *A. baumannii* study. Finally, in two studies aiming to identify mechanisms of resistance to polymyxin antibiotics in *A. baumannii*, IS-mediated gene disruption events were identified by ISMapper in multiple isolates whose resistance phenotypes could not otherwise be explained[277].

3.2 Reconstructing AMR elements in *Salmonella* Kentucky

3.2.1 Multi-drug resistant *S.* Kentucky ST198

S. Kentucky can cause gastroenteritis in humans but is most often isolated from non-human sources, most commonly from poultry[278,279]. This host preference has been assisted by

S. Kentucky gaining virulence plasmids from *E. coli*, which have been shown to be important for the colonisation of poultry, such as the virulence plasmid ColV[280]. *S.* Kentucky is able to colonise the poultry gastrointestinal tract and is transmitted via the faecal-oral route[281]. *S.* Kentucky has been shown to cause community-based outbreaks, many of which have occurred after travellers have returned to their home countries from Africa[279].

The majority of MDR *S.* Kentucky isolates cluster in a single sequence type (ST), ST198[279]. Before the turn of the century, *S.* Kentucky was susceptible to all antibiotics. Since the turn of the century, MDR has begun to emerge[283]. Ciprofloxacin resistant *S.* Kentucky first surfaced in 2002, when it was noted that *S.* Kentucky isolated from French travellers returning from Africa displayed resistance to this fluoroquinolone-based drug[283,284]. *S.* Kentucky isolates resistant to one or more antibiotics (namely; amoxicillin, gentamicin, nalidixic acid, sulfonamides, tetracycline and streptomycin) have also been isolated between 2000 and 2005[283,285]. AMR doubled amongst *S.* Kentucky isolates located in France between 2000-2008 and 2009-2011, including resistance to fluoroquinolones[286]. In Canada, resistance to more than one antibiotic in *S.* Kentucky increased and, by 2013, the majority of *S.* Kentucky isolates found in North America were MDR[287,288]. In each of these studies the majority of isolates, and all MDR isolates, belonged to ST198. This suggests that the recent global spread of *S.* Kentucky may reflect the expansion of a single clone, driven by the emergence of antimicrobial resistance.

3.2.1.1 MDR in *S.* Kentucky is caused by the SGI

MDR in *S.* Kentucky is mostly attributed to the acquisition of the *Salmonella* genomic island (SGI) into the chromosome[282]. The SGI is a 43 kbp element first characterised in *S. enterica* serovar Typhimurium strain DT104[38], and situated between the *trmE* and *yidY* genes on the chromosome[289]. In general, the SGI consists of a 27.4 kbp backbone, with a variable region that contains a 15 kbp complex class I integron (In104), inserted next to the resolvase gene (*resG*/*S027*) of the backbone. This class I integron contains an antibiotic resistance region and is flanked by a 5 bp duplication[290]. The SGI itself is flanked by 18 bp imperfect repeats, which assist with excision and the formation of a circular extrachromosomal structure. This structure is not self-transferrable but is able to excise itself using the integrase (*int*) gene with the help of an IncA/C plasmid[291–294]. The SGI integrates into the chromosome in a site-specific manner, and has been shown to integrate into *E. coli* at the 3' end of its *trmE* gene[291]. The structure itself seems to be unstable in the genome when the selective pressure of antibiotics is absent[39].

Several variants of the SGI have been defined, each labelled using letters of the alphabet (currently SGI1-A to SGI1-V)[290]. These variants usually differ in the composition of the integron, and each variant contains different antibiotic resistance genes. One variant of the SGI, known as SGI2, differs not only in the composition of the integron, but also in the site at which the integron is inserted into the SGI backbone[295]. Four main types of SGI have so far been described in *S.* Kentucky by Le Hello *et al.*[286] – SGI1-K, SGI1-P, SGI1-Q and SGI2 (previously known as SGI1-J). The SGI1 variants are related to the SGI1-H variant found in *S. enterica* serovar Newport as they differ slightly in their backbone, compared to other SGI1 types[296]. There has been a deletion between *S005* and *S009* due to the introduction of the the insertion sequence IS*1359*, resulting in truncated *S005* and *S009* genes[170]. All three SGI1 variants found in these *S.* Kentucky isolates also contain IS*26*, which truncates *S044*, the final gene of the backbone[282]. SGI1-K contains a mercuric chloride resistance region (*mer*) before In104[296]. The tetracycline resistance in SGI1-K is due to *tet(A)*, ampicillin resistance is due to a transposon containing the bla_{TEM} gene, and streptomycin resistance is caused by a transposon containing *strAB*[296]. SGI1-P and SGI1-Q don't contain the *mer* resistance region or In104 , instead SGI1-P contains only the transposase carrying bla_{TEM}, and SGI1-Q has no resistance genes present in the island at all (Simon Le Hello, Institut Pasteur, personal communication).

3.2.2 Genomic study of *S.* Kentucky ST198

The analysis presented below is my contribution to the larger study with collaborators François Xavier-Weill and Simon Le Hello at the Institut Pasteur, Paris. The specific aims of the project were to understand:

i) when and where MDR *S.* Kentucky ST198 evolved;

ii) what role the SGI played in its evolution; and

iii) how the SGI changed within this clone over time.

3.2.3 Methods

3.2.3.1 Isolate collection and AMR phenotyping

The isolates examined by Le Hello *et. al*[286] as well as additional ST198 isolates collected by the Institut Pasteur were used in this study. A total of 88 *S.* Kentucky ST198 genomes were sequenced at GATC Biotech (Germany) on the Illumina HiSeq platform, and the data sent to Melbourne for analysis. Antimicrobial susceptibility profiling was performed on all isolates, using either disk diffusion breakpoint or broth microdilution by Bio-Rad in Marnes-La-Coquette, France.

3.2.3.2 Mapping and phylogeographic analysis

Short reads from each isolate were mapped to the reference genome 1922 using the mapping pipeline RedDog v1b4 (https://github.com/katholt/RedDog) to identify SNPs. RedDog uses Bowtie2 v2.2.3[297] with the sensitive local method and a maximum insert size of 2000 to map all genomes to the reference genome. SNPs were then identified using SAMtools v0.0.19[266], and alleles at each locus were determined by comparing to the consensus base in that genome, using SAMtools pileup to remove low quality alleles (base quality <= 20, read depth <=5 or a heterozygous base call). SNPs were filtered to exclude those present in repeat regions, phage regions or the SGI. Gubbins v1[298] was run using default settings to identify and remove SNPs in recombinant regions. The final SNP set consisted of 1,144 SNPs.

To estimate a Bayesian phylogeny with divergence dates, an alignment of SNP alleles was passed to BEAST v1.7.5[299], in addition to isolation dates for each genome. The model parameters were as follows: GTR+Γ substitution model, lognormal relaxed clock, constant population size. Ten independent BEAST runs of 50 million iterations were combined, representing 495 million Markov chain Monte Carlo (MCMC) generations after burn-in removed. Parameter estimates were calculated using Tracer v1.6[300]. A maximum clade credibility tree was generated using TreeAnnotator v1.7.5[200]. To test the robustness of the molecular clock signal, five further BEAST runs with randomised tip dates were generated using the same model (Figure 3.1). Ancestral state reconstruction was performed with BEAST, with geographic regions determined by the United Nations subregion geoschemes[301], modeled as discrete states.

3.2.3.3 Assembly and annotation

All reads were filtered using FastXToolKit v0.0.14[302] to remove all reads containing N's, and Trimmomatic v0.30[303] was used to remove any reads with an average phred quality score below 30. Each isolate was assembled using SPAdes v3.5[304] using a kmer range of 21, 33, 55, 65 and 75. Scaffolding was performed using SSPACE v3.0[305] and GapFiller v1.10[306] with default settings. All assemblies were ordered against the 1922 reference genome using Abacas v1.3.1[307]. Prokka v1.10[308] was used to annotate each assembly using a preferential protein database made up of coding sequences from the 1922 reference genome, the ARGAnnot resistance database, the SGI1, SGI1-K and SGI2 references (accessions AF261825, AY463797 and AY963803).

3.2.3.4 Reconstructing the SGI sequences

ISMapper and the assembly graph viewer Bandage[309] were used to piece together segments of the SGI. To do this, each assembly was queried with BLAST to identify which contigs contained SGI backbone and antibiotic resistance genes. Each assembly was also queried for IS*26* using ISMapper's assembly improvement mode, identifying contigs that contained IS*26* flanking sequence. Contigs containing flanking IS*26* sequence with SGI genes or antibiotic resistance genes were hypothesised to be part of the SGI. Both pieces of information (BLAST and ISMapper) were used in conjunction with the reference SGI1-K sequence to determine which contigs could be joined together. In some cases, it was unclear whether IS*26* flanked resistance genes were part of the SGI or a plasmid. In these cases Bandage was used to examine the assembly graphs and determine the paths linking the SGI, IS*26* and the resistance genes, providing additional evidence for contig connection. IS*26* copy number was estimated using assembly improvement mode in ISMapper, by counting the total number of contigs pairs with either left or right flanking sequences.

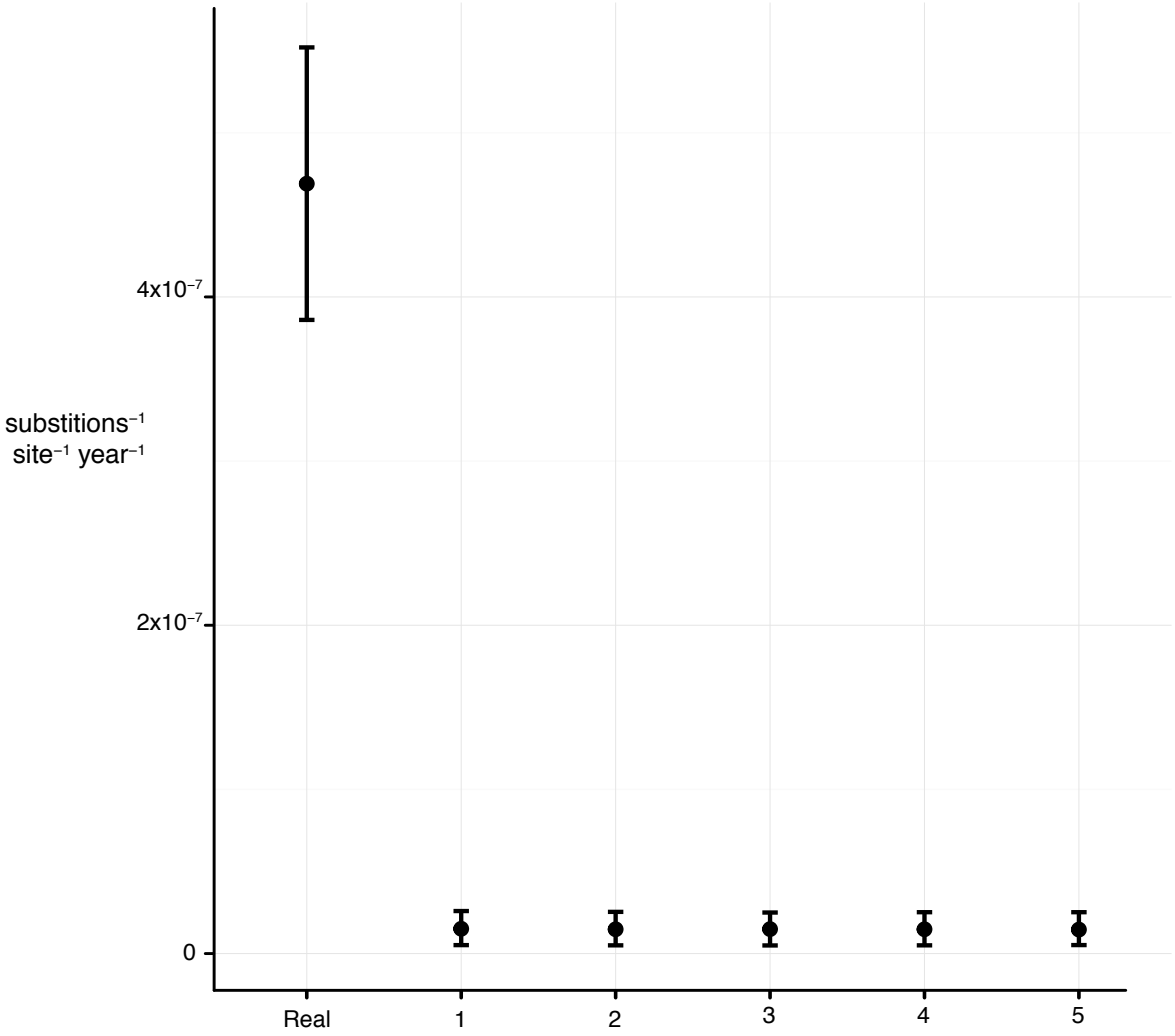

Figure 3.1: Mutation rate estimates for real and randomised tip dates in *S.* Kentucky. First column, real mutation rate. Subsequent columns show mutation rate when tip dates are randomised. Black circles are the mean rate estimated by BEAST, with error bars showing 95% highest posterior density (HPD).

3.2.4 Phylogeographic analysis of *S.* Kentucky ST198 based on whole genome SNP data

In total, 1,144 SNPs were identified within the core genome of ST198. The alignment of these SNPs and the years of isolation were used to construct a dated phylogenetic tree using BEAST (Figure 3.2). From the BEAST analysis, it is estimated that the MDR sub-clone of ST198 emerged around 1992 (95% HPD, 1988-1995) in Egypt. The SGI1 was acquired in the early 1990s, in the common ancestor of the ST198 sub-clone (arrow, Figure 3.2). These results suggest that after the acquisition of SGI1, there was a single clonal expansion of ST198. The resulting sub-clone had an MDR phenotype.

In addition to SGI1, the ST198 MDR clone rapidly accumulated base substitution mutations in *gyrA* (DNA gyrase) and *parC* (DNA topoisomerase), causing additional resistance to ciprofloxacin and naladixic acid. The first mutation occurred in *gyrA* codon 83 (TCC -> TTC) (circa 1993, light purple, Figure 3.2), and was then followed by a mutation in codon 80 of *parC* (AGC -> ATC), around 1997 (pink, Figure 3.2). This accompanied a dramatic clonal expansion, with the clone spreading from Egypt into other geographical locations. During this expansion, other mutations arose in codon 87 of *gyrA* (3 independent mutations to GGC, AAC and TAC, dark purple shades, Figure 3.2). Multiple independent transfers of ST198 out of Egypt and Northern Africa are evident, with two clades, carrying either of the TAC and AAC *gyrA* codon 87 mutations; the former spread into East Africa, Middle Africa, Southern Asia, Europe and Western Asia; the latter spread to South-Eastern Asia, Europe and West Africa (Figure 3.2).

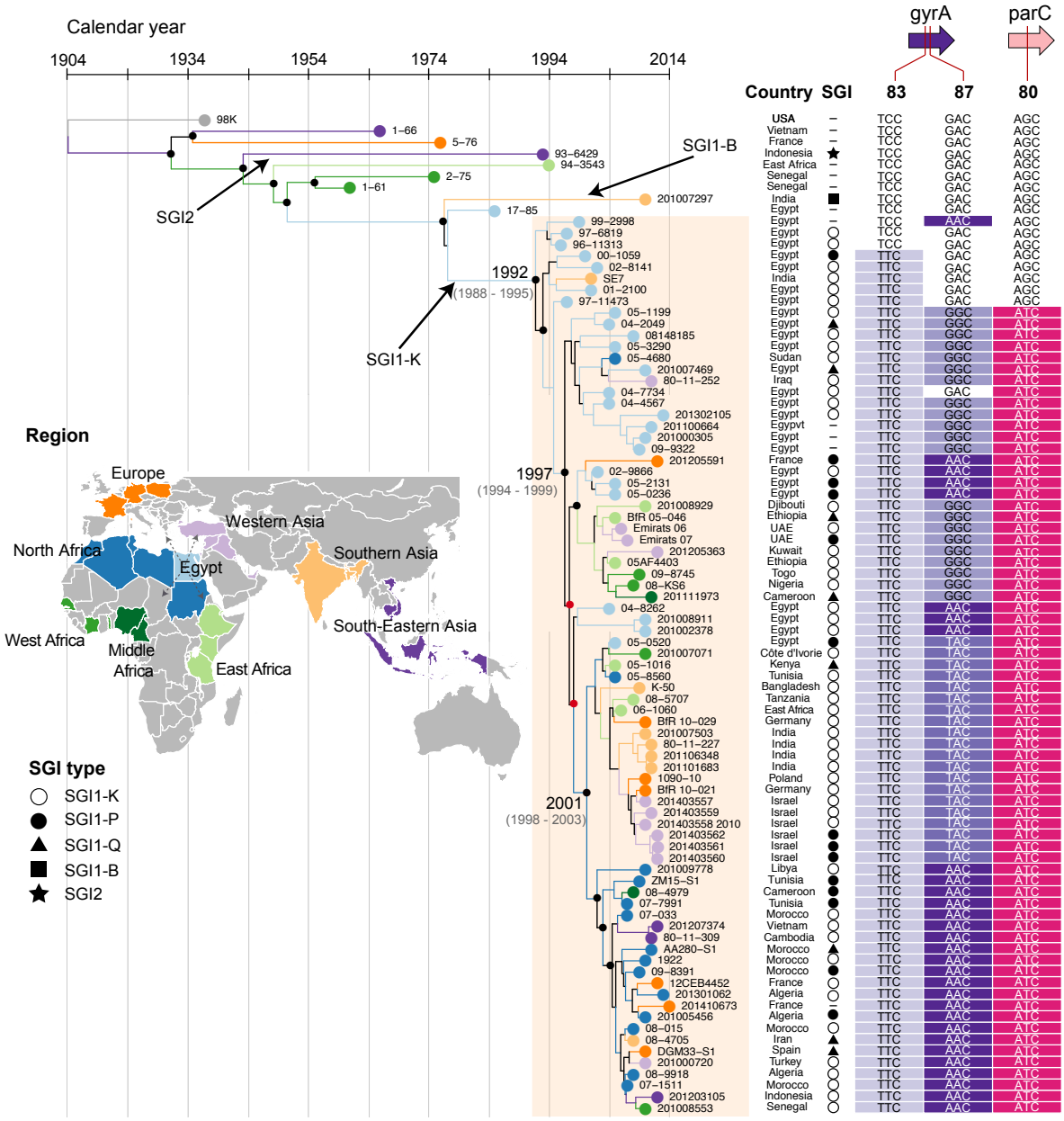

Figure 3.2: Phylogeographic reconstruction of *S* Kentucky ST198. Phylogeny is maximum clade credibility tree estimated from BEAST. Nodes and branches are coloured by region as per map. SGI type in each isolate indicated by symbol as per legend. Arrows indicate branches where different SGI types were acquired. Mutations in codons 83 and 87 in *gyrA* shown in purple, and mutations in codon 80 of *parC* shown in pink.

3.2.5 Variation within the SGI in *S.* Kentucky ST198

Presence of any SGI backbone genes was taken as evidence of the presence of the SGI (Figure 3.3). Some isolates had large deletions of the backbone (eg: deletions from *S011* to *S026*, or *int* to *S026*, Figure 3.3), but still contained the MDR region between *trmE* and *yidY* (Figure 3.4)

Almost every isolate in this study was found to have a distinct SGI structure (Figure 3.4). In addition to large deletions of the SGI backbone, some isolates had inversions of whole or part of the resistance gene segment of the island, with various deletions and rearrangements of the AMR transposons (Figure 3.4). There were also multiple different IS*26* insertion sites within the resistance elements of the island (Figure 3.4). In the cases where the SGI1 subtype detected was SGI1-P or SGI1-Q (containing few or no resistance genes), some ST198 isolates had imported large resistance plasmids instead that provide an MDR phenotype (Figure 3.5). This was also the case in isolates where the SGI had been partially deleted or deleted in its entirety.

In the isolates in this study, the SGI and associated resistance genes were spread across an average of three contigs, due to multiple insertions of IS*26* within the chromosome. IS*26* is 820 bp long and encodes a single transposase, with 14 bp terminal repeats on each end[310]. Each of these three SGI1 subtypes contained one or more copies of IS*26*. All genomes in the ST198 MDR sub-clone have copies of IS*26*, with no genomes outside of this clade containing IS*26* (Figure 3.3). IS*26* copy number varies between isolates, with some isolates having as many as 12 copies of IS*26*. Not all of these IS*26* sites are within the SGI. Once IS*26* has become part of the chromosome, in some cases it has replicated throughout the genome of the isolate (Figure 3.6). There appears to be no relationship between SGI subtype and IS*26* copy number, with IS*26* varying in copy number across the entire sub-clone (Figure 3.6).

The recently described mechanism used by IS*26* to tranpose may provide an explanation as to why the SGI variants in these isolates are so dynamic. During transposition, IS*26* extracts itself from the donor DNA molecule, as well as DNA lying upstream of it between itself and another IS*26* element, and uses this to form a translocatable unit[311]. It then finds another IS*26* element in the receiving DNA molecule, and inserts itself, as well as the excised donor DNA next to it, forming a tandem array of IS*26*s in direct orientation[311]. Using this model, figures 3.4 and 3.3 illustrate that IS*26* is likely the causative agent for many of the deletions, inversions, and transpositions within the SGI. It may also be inferred that IS*26* was responsible for the main variants of SGI1 (SGI1-K, SGI1-P and SGI1-Q) seen in this dataset.

These results demonstrate that the SGI in *S.* Kentucky ST198 is extremely dynamic. The whole genome data shows that there is no conservation of any of these three subtypes across the tree – every subgroup contained a mix of SGI types with no clear pattern. By using ISMapper to search for IS*26* elements within the assemblies of these genomes, it was possible to reconstruct a complicated resistance region using only short read data.

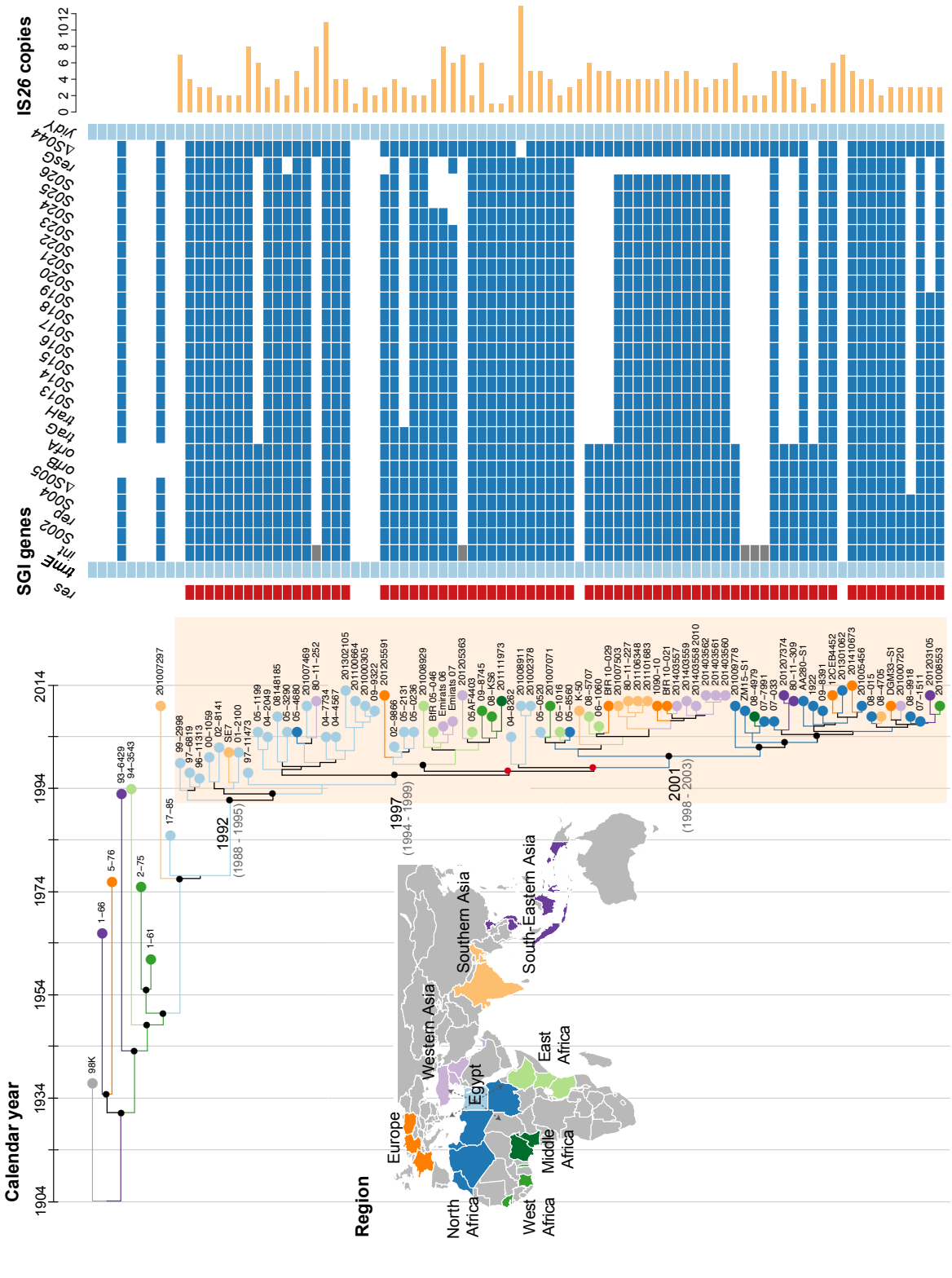

Figure 3.3: Presence of SGI backbone genes and IS26 copy number in each *S. Kentucky* genome. Left, maximum clade credibility tree estimated from BEAST, with branches and tips coloured by region as per map. Country of isolation is listed next to each genome. Presence of SGI genes are shown in heatmap; red, resistance region; light blue, flanking chromosomal genes (*trmE* and *yidY*); dark blue, SGI backbone genes; grey, partial gene deletion. Orange bar chart shows IS26 copy number for each genome, estimated by ISMapper.

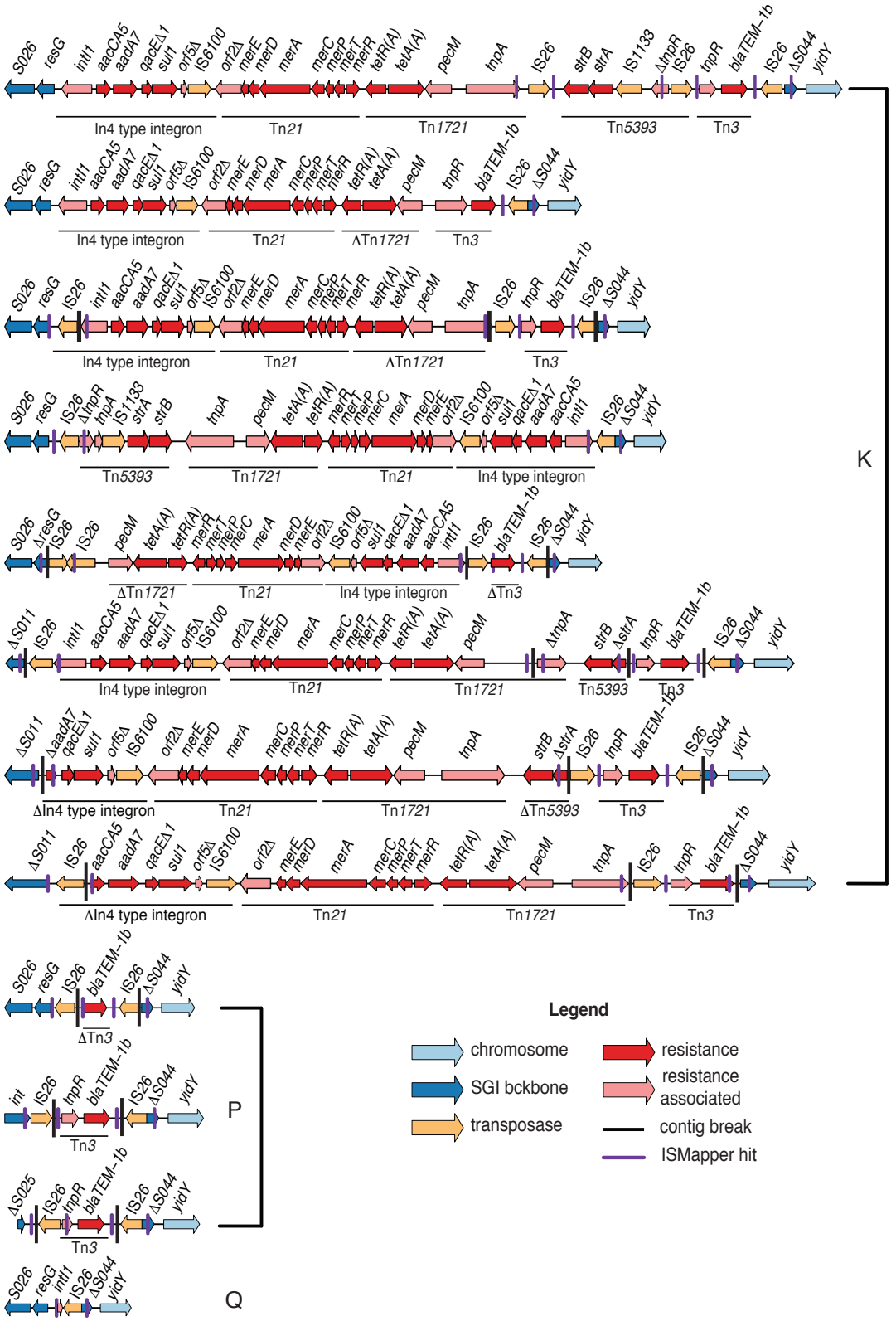

Figure 3.4: Examples of the variation found with each SGI1 type found in *S.* Kentucky. Genes are represented by arrows and coloured by type as per legend. Contig breaks are shown by black bars, ISMapper hits by purple bars.

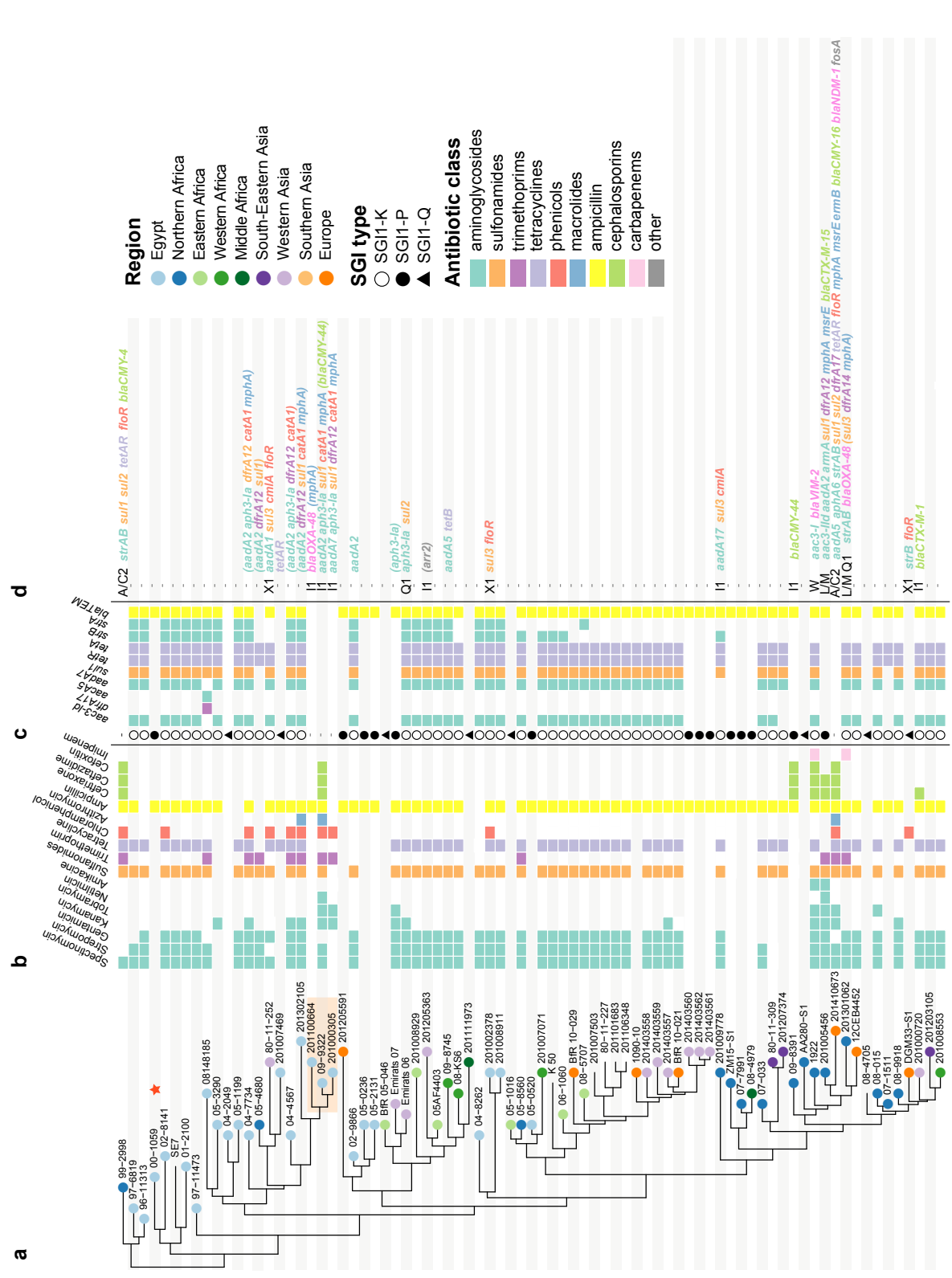

Figure 3.5: Presence of antibiotic resistance genes within each genome of S. Kentucky. a, Phylogenetic tree of ST198 MDR clone with tips coloured by region. **b,** Resistance phenotype with drugs coloured by class as per legend. **c,** Symbols represent SGI type in isolate, as per legend. Resistance genes present on the SGI, coloured by the class of antibiotic they confer resistance to. **d,** Plasmid replicon detected in each isolate, with resistance genes listed that are present on that plasmid. Gene names in brackets indicate unknown location. Gene names are coloured by antibiotic class they confer resistance to.

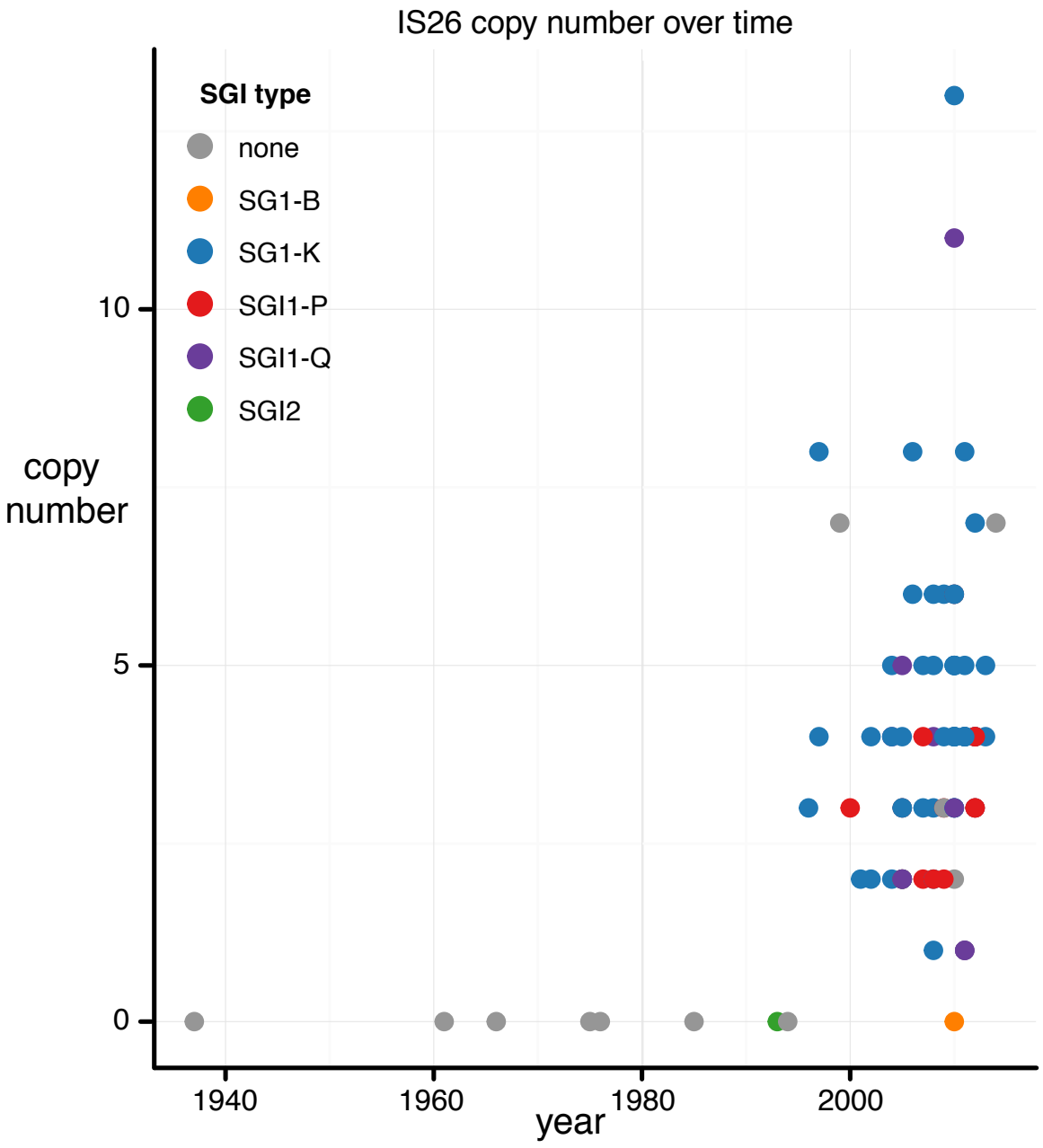

Figure 3.6: IS*26* copy number in each genome. *x*-axis, year genome isolated; *y*-axis, IS*26* copy number. Points are coloured by SGI type, as per legend.

3.3 The spread of antibiotic resistance in *S.* Typhi

3.3.1 Multi-drug resistant *S.* Typhi

S. Typhi is an invasive, human-adapted pathogen that causes typhoid fever[312]. *S.* Typhi causes a large burden of disease in the developing world, especially in Asia and Africa[313,314]. Whilst there are vaccines for *S.* Typhi, they are only moderately effective[315]. The best control measure to date are antibiotics. However, resistance to antibiotics emerged in the 1950s[316], and by the 1980s the majority of *S.* Typhi was MDR (resistant to chloramphenicol, ampicillin and trimethoprim-sulfamethoxazole)[317]. MDR genes in *S.* Typhi are typically are carried on large IncHI1 plasmids[318]. Initially, there were many different types of IncHI1 plasmids circulating within the global *S.* Typhi population in multiple different chromosomal backgrounds[52]. However, since 1995, the majority of the MDR typhoid has been caused by isolates of a single, globally disseminated, clone called H58, carrying a single IncHI1 plasmid MLST type (PST6) encoding all the MDR genes[319]. In this plasmid, the MDR genes are typically present on a Tn*2670*-like complex transposon flanked by copies of IS*1*, which are believed to mobilise the element[51,101].

3.3.2 Genomic analysis of a global collection of *S.* Typhi

To investigate the global population structure of *S.* Typhi and the evolution of MDR, members of the International Typhoid Consortium collaborated to WGS 1,832 isolates originating from 63 countries[276]. Phylogenetic analysis revealed that 853 isolates (47%) belonged to the H58 clone, of which 63% were MDR. 61% of MDR H58 genomes carried all seven Tn*2670* genes (Figure 3.8), and in 71% of cases these genes were present on the IncHI1 PST6 plasmid. The remaining 149 isolates carried the typical MDR transposon genes without evidence of the IncHI1 plasmid, or any other plasmids, being present. It was therefore hypothesised that the MDR transposon may have been mobilised to the chromosome via the flanking copies of IS*1*. This hypothesis was investigated using ISMapper to interrogate the *S.* Typhi genomes.

3.3.3 ISMapper analysis

The ISMapper analysis presented here is my contribution to the larger study, which was published in Wong et. al., *Phylogeographical analysis of the dominant multidrug-resistant H58 clade of* Salmonella *Typhi identifies inter- and intracontinental transmission events.* 2015, Nature Genetics *47*, 632-639.

In order to determine the location of the MDR transposon, which is flanked and mobilised by IS*1*, all H58 isolates were screened for IS*1* using ISMapper's typing mode to identify insertions relative to the reference chromosome *S.* Typhi CT18 (accession NC_003198). ISMapper was able to quickly screen hundreds of *S.* Typhi genomes, and in total found four novel IS*1* positions that were possible locations for the insertion of the MDR transposon. Almost all isolates lacking the IncHI1 MDR plasmid contained an IS*1* insertion site near *cyaA*, between STY3618 and STY3619 (Figure 3.7). Long-read sequencing (conducted at the Sanger Institute) confirmed that this was the location of the MDR transposon - it had moved from the plasmid into the chromosome (Figure 3.8). This insertion site in *cyaA* was not conserved amongst all isolates, with 109 genomes containing the MDR transposon at this location. There were three other transposon insertion sites : *yidA* (n=25), *fbp* (n=1) and STY4438 (n=9) (Figure 3.7). Long read sequencing confirmed the sites at *yidA*, and PCR (conducted by Derek Pickard at the Sanger Institute) confirmed the remaining two sites[276].

3.3.4 Conclusions

Tracking the movement of antibiotic resistance transposons from plasmids to the chromosome is important, as elements on the chromosome are much more stable than plasmids[320]. Chromosomal elements are less likely to be lost in the absence of selective pressure from exposure to antibiotics[320]. In this study, the movement of the MDR transposon from the plasmid to the chromosome was shown to have occurred on multiple independent occasions. These results suggest the strong selective pressure within *S.* Typhi to maintain an MDR phenotype without the additional burden of replicating a large plasmid.

This study illustrates the importance of being able to use a high throughput method to detect insertion sites within large, short read datasets. It is also an additional validation of ISMapper, as each insertion site detected by ISMapper was independently confirmed using either long-read

sequencing or PCR. In this case, hundreds of existing read sets could be quickly screened for IS*1* to determine the location of the transposon, which is more cost effective and feasible than running additional assays such as long-read sequencing or PCR.

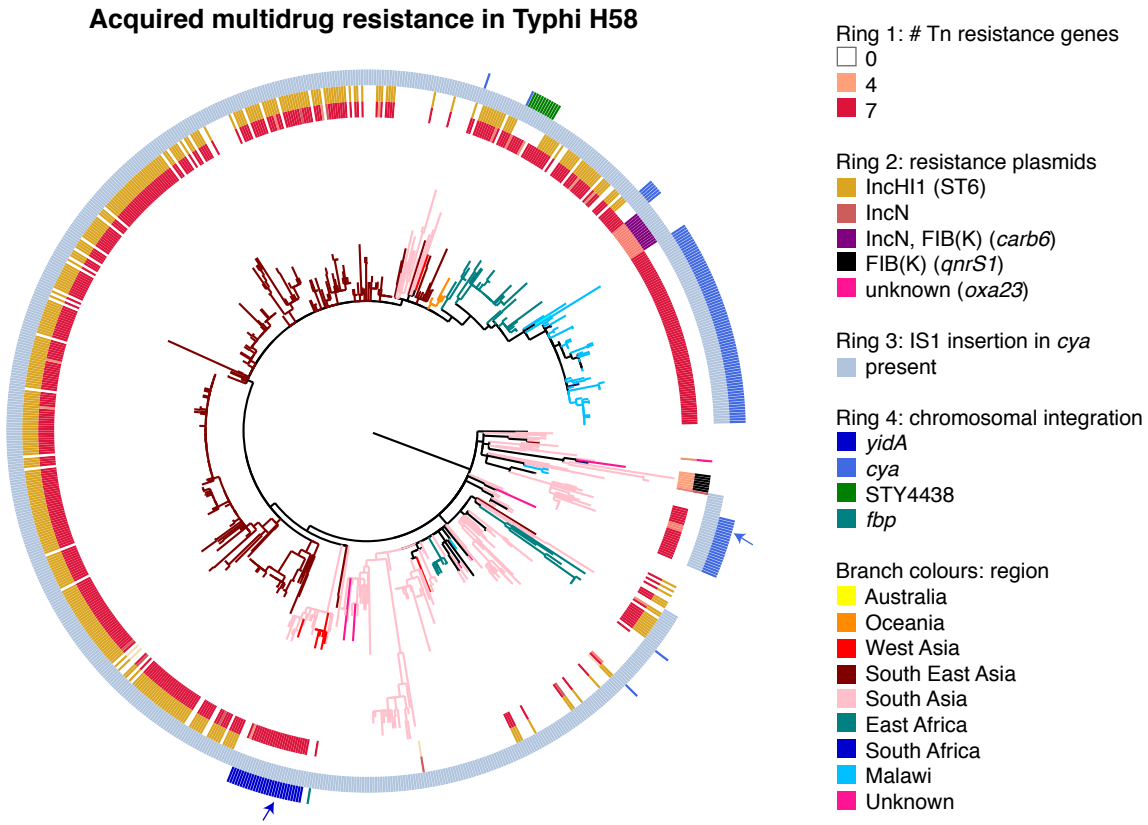

Figure 3.7: Phylogeny of H58 *S.* Typhi isolates. Branches coloured by region. Innermost ring, red, shows number of resistance genes found on the Tn*2670*. Second ring shows plasmid replicon type for each isolate. Third ring shows presence of IS*1* in *cyaA* and outer ring shows Tn*2670* insertion site. Figure adapted from Wong et. al., 2015.[276]

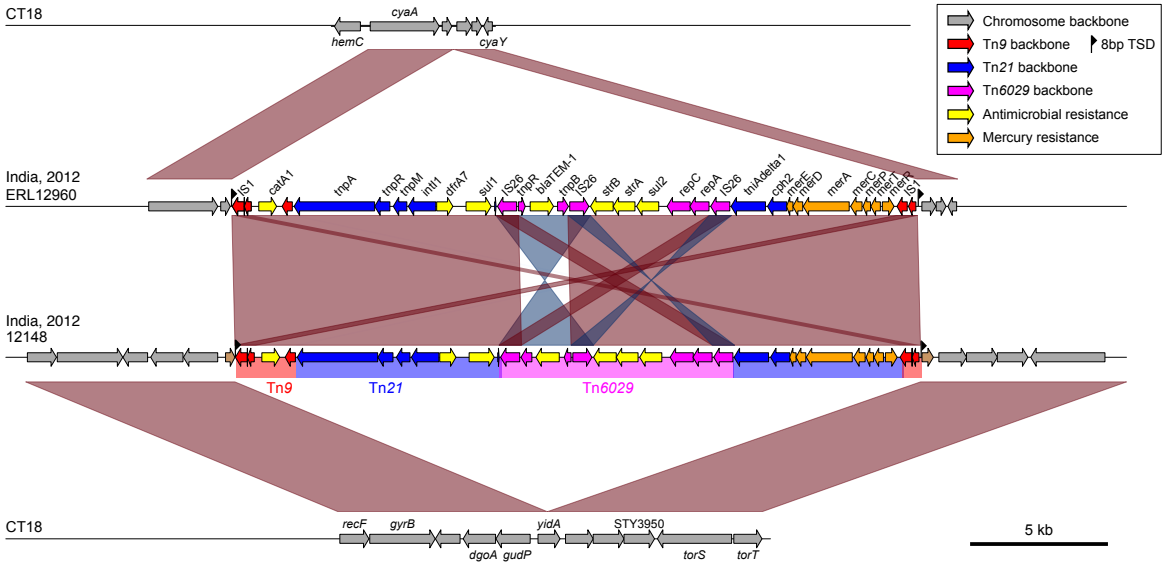

Figure 3.8: Location of transposon insertion sites in *cyaA* and *yidA*. Top, CT18 reference at *cyaA*. Next two panels show transposon inserted in *cyaA* and then *yidA*. Bottom, CT18 reference at *yidA* site. Figure reproduced from Wong et. al., 2015.[276]

3.4 Antibiotic resistance in *Acinetobacter baumannii*

3.4.1 *A. baumannii* and resistance

A. baumannii is becoming a pathogen of international importance, especially within hospitals[321]. It is one of the ESKAPE pathogens (*Enterococcus faecium*, *S. aureus*, *K. pneumoniae*, *A. baumannii*, *Pseudomonas aeruginosa* and *Enterobacter spp.*), identified by the US Infectious Diseases Society as the most concerning bacteria escaping antibiotic treatment, as it is becoming increasingly drug resistant[322]. The typical *A. baumannii* chromosome includes intrinsic genes encoding beta-lactamases and efflux pumps, which are able to pump antibiotics out of the cell. These mechanisms, together with horizontally acquired resistance genes, can combine to create isolates that are resistant to a very broad spectrum of antibiotics.

A. baumannii have intrinsic resistance to chloramphenicol and florfenicol due to the capsule surrounding the cell[323]. Further AMR to first line drugs arose in the 1980s through the chromosomal acquisition of a large genomic island known as AbaR within Global Clone 1 (GC1) and Global Clone 2 (GC2). During the 1990s, resistance to fluoroquinolones arose,

leaving few treatment options[324]. More recently, AMR to third-generation cephalosporins has emerged, primarily driven by IS upregulation of *ampC*[325]. Carbapenem resistance is now on the rise, leaving colistin (of the polymyxin class of antibiotics) and tigecycline as last-line treatment in these cases, despite the fact that colistin has been reported to be toxic[326].

The following two studies demonstrate the usefulness of ISMapper in detecting complex resistance mechanisms to last resort antibiotics in *A. baumannii*.

3.4.2 IS-mediated resistance in *A. baumannii* isolated from a Vietnamese intensive care unit

The ISMapper analysis presented here is my contribution to a larger study, published in Shultz *et. al.*, *Repeated local emergence of carbapenem resistant* Acinetobacter baumannii *in a single hospital ward*, 2016, Microbial Genomics.

A study of ventilator associated pneumonia was conducted in an intensive care unit in Ho Chi Minh City hospital in Vietnam between 2003 and 2012[69]. The study consisted of 147 *A. baumannii* isolates from patients with ventilator associated pneumoniae. Isolates from 2003 - 2007 were sensitive to the carbapenem, imipenem, and isolates from 2008 - 2012 were imipenem resistant. Each isolate was sequenced on the Illumina HiSeq 2500, generating 100 bp paired end reads. Based on earlier results of imipenem resistant isolates carrying *oxa23* within the hospital[327], it was hypothesised that there would be a dominant *A. baumannii* imipenem resistant clone circulating within the hospital.

In this study, the imipenem resistance was found to be due to one of four carbapenemase resistance genes; *oxa23*, *oxa58*, *oxa72* or *ndm-1*. Resistance caused by *oxa58*, *oxa72* and *ndm-1* was mostly plasmid associated and rare. However, resistance caused by *oxa23* was the most common, present in 93% of imipenem resistant isolates, and was located within either an Tn*2006* or Tn*2008*-like transposon, which are mobilised by IS. The majority of the imipenem resistant isolates (82.5%) belonged to the GC2 clone, and all were carrying the *oxa23* gene.

Phylogenetic analysis of the GC2 group revealed that the isolates split into two distinct lineages that diverged in approximately 2002. Lineage 1 contained only imipenem sensitive isolates, whilst lineage 2 contained mostly imipenem resistant isolates falling into five distinct sub-clones (Figure 3.9). It was hypothesised that each of these imipenem resistance sub-clones

had independently evolved through the acquisition of ISAba1-associated transposons Tn*2006* or Tn*2008* due to their phylogenetic separation. The aim of this analysis was to investigate the evolution of these sub-clones with ISMapper.

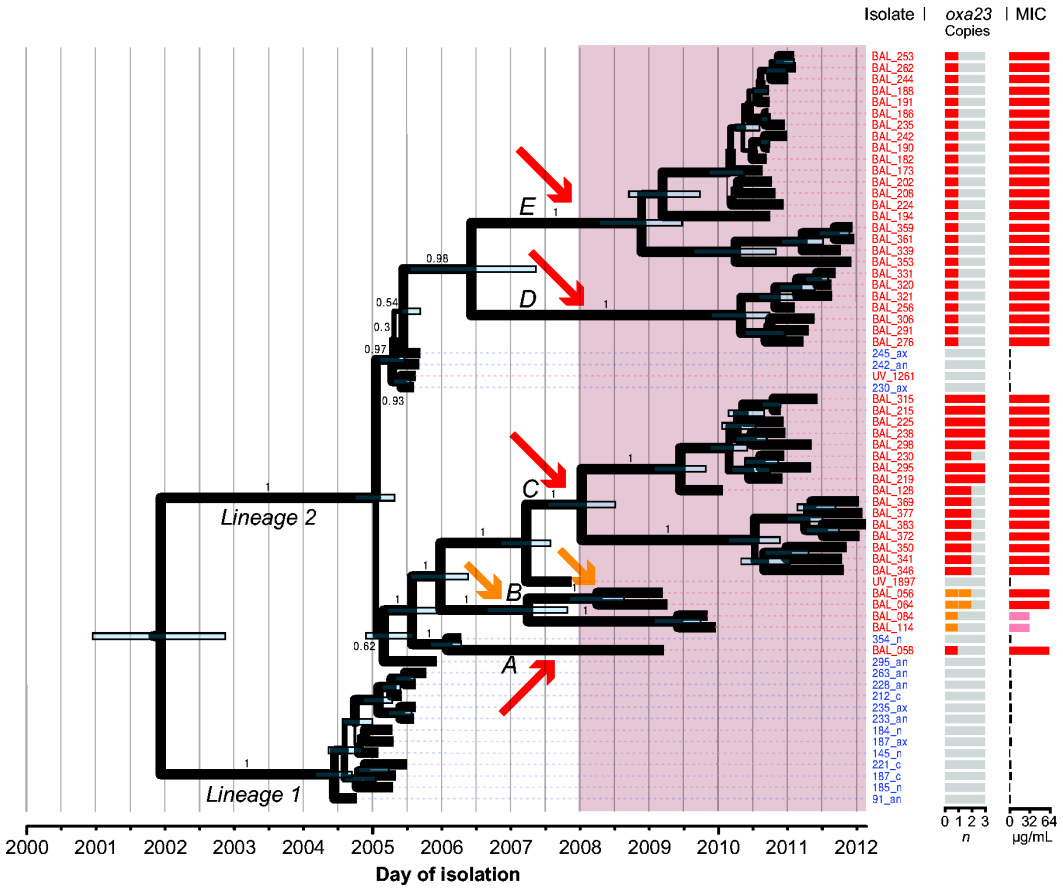

Figure 3.9: Core genome phylogeny for GC2 *A. baumannii* in the ICU. BEAST maximum clade credibility tree; shading indicates the period during which imipenem was used for the empirical treatment of VAP in the ICU. Isolate labels are coloured to indicate source: red, VAP; blue, asymptomatic carriage. Node bars indicate 95 % HPDs for divergence dates; node labels and branch line thickness indicate posterior support. The two main lineages (1 and 2) and five imipenem-resistant subclades (A–E) are labelled, arrows indicate inferred *oxa23* carbapenemase acquisition events: red, Tn*2006*; orange, Tn2008VAR. *Oxa23* gene copy number and MICs for imipenem are indicated on the right. Figure reproduced from Shultz et. al., 2016.[69]

3.4.2.1 ISMapper analysis

To detect the ISAba1 positions in the GC2 genomes, ISMapper was used in typing mode with GC2 strain 1656-2 reference genome (accession CP001921.1). There were multiple copies of

ISAba1 within the genome, however each imipenem resistant sub-clone harboured at least one unique ISAba1 insertion (arrows, Figure 3.9). These unique sites were consistent with independent acquisitions of one of the *oxa23* carrying transposons within each subclone. The majority of isolates were also resistant to ceftriaxone and other third generation cephalosporins. In most cases there were no horizontally acquired AMR-associated genes that could explain this phenotype. The ISMapper analysis detected the insertion of ISAba1 upstream of the beta-lactamase gene *ampC*. Insertions of ISAba1 or ISAba125 upstream of *ampC* have been shown to cause high level resistance to third generation cephalosporins. This resistance is caused by overexpressing *ampC*, resulting in the destruction of the cephalosporin molecules, producing a carbapenem-resistant phenotype[130,325]. This demonstrates the utility of ISMapper to identify genetic mechanisms for unexplained antibiotic resistance phenotypes.

3.4.3 IS-mediated resistance to polymyxins in *A. baumannii*

Two types of polymyxins are commonly used to treat *A. baumannii* and other highly drug-resistant pathogens - polymyxin B and polymyxin E (also known as colistin). Polymyxins interact with lipid A in the lipopolysaccharide (LPS) on the surface of the cell, disrupting it to cause cell death[328]. Modifications to the lipid structure surrounding the cell prevents this interaction and renders the cell resistant. SNP mutations in the two-component regulators *pmrAB* or *phoPQ* confer resistance by altering the LPS to prevent polymyxin binding[329,330]. ISAba1 insertions within either *lpxA* or *lpxC* are also known mechanisms of polymyxin resistance[331]. Recently, the first horizontally transferred polymyxin resistance gene, *mcr-1*, was described[332]. However the majority of clinical resistance to colistin is associated with mutational resistance in core genes.

Resistance to polymyxins has recently emerged in *A. baumannii*[333]. Polymyxin antibiotics are the last line of defense against *A. baumannii* infection[334], so it is imperative to understand mechanisms of resistance. In the following two studies, resistance to polymyxin B and colistin in *A. baumannii* was investigated *in vitro*, demonstrating a significant role for IS in the evolution of resistance.

3.4.3.1 Polymyxin B resistance in Singapore *A. baumannii* isolates

The ISMapper analysis presented in this study is my contribution to a larger work published in Lim *et. al.*, *Multiple genetic mutations associated with polymyxin resistance in* A. baumannii, 2015, Antimicrobial Agents and Chemotherapy 59(12), 7899-7902.

3.4.3.1.1 Introduction to study

In this study, ten clinical *A. baumannii* isolates susceptible to polymyxin were obtained by Li Yang Hsu's group in Singapore. Two of these isolates (1 and 2) were taken from patients that were later treated with polymyxin and developed resistance. The remaining eight susceptible isolates were passaged through polymyxin *in vitro* to generate isogenic resistant mutants. All ten resistant-susceptible pairs were sequenced on the Illumina HiSeq. Li Yang Hsu's group performed a SNP analysis to identify SNPs and indels that were responsible for the polymyxin resistance developed through either (i) treatment with polymyxin or; (ii) passaging through polymyxin. The SNP analysis identified mutations in *pmrB*, *lpxD*, *lpxA* and *lpxC* in six of the ten resistant isolates (isolates 1, 2, 6, 7, 9 and 10, Table 3.1). Additionally, one isolate, 2, contained many more SNPs that than the others. The remaining four polymyxin resistance phenotypes were unable to be explained by the SNP analysis (isolates 3, 4, 5 and 8, Table 3.1).

3.4.3.1.2 ISMapper analysis

ISMapper was used in typing mode to screen all 20 isolates for ISAba1 and IS*15* insertions against the *A. baumannii* reference A1 (accession CP010781). Insertions were identified in *lpxA* in isolates 4 and 5. Isolate 3 contained an ISAba1 insertion within *lpxC*. The ISMapper analysis did not identify any IS insertions within isolate 8 that could explain the polymyxin resistance. Closer inspection of this isolate's assembly with the assembly graph viewer, Bandage, found that the *lpxC* gene had been interrupted by insertion of the ISAba1-flanked genomic resistance island, AbaR. ISMapper had been unable to detect this insertion as the insertion of AbaR had mediated a large genomic rearrangement, which was difficult to detect using a mapping approach. In addition, isolate 2, which had been noted to contain a high number of SNPs, was found to have an IS*15* insertion within *mutS*, a DNA mismatch repair gene. The disruption of *mutS* may be linked to the high number of SNPs detected, as disruption of this gene has previously been shown to create a hypermutator phenotype[335,336].

3.4.3.2 Colistin resistance in Ho Chi Minh City *A. baumannii* isolates

3.4.3.2.1 Introduction to study

In this study, Stephen Baker's group at the Oxford University Clinical Research Unit (OUCRU) in Vietnam, and Sanger Institute, investigated colistin resistance in *A. baumannii*. Two independent cultures of the same strain were serially passaged through different concentrations of colistin, until they were able to grow at a colistin concentration of 128 mg/L. Colonies were selected for downstream analysis at four different colistin concentrations: 0 mg/L, 32 mg/L, 64 mg/L and 128 mg/L. Each colony was then inoculated into two separate broths, one containing the same concentration of colistin and one without colistin, generating a pair of isolates (A and B). Isolates A and B at each concentration were sequenced on the Illumina platform. Stephen Baker's group performed a SNP analysis on all of the isolates. Two isolates contained SNPs within the *lpxA* gene (45A and 45B), but the remaining isolates did not have any causative SNPs that could explain their resistance to colistin.

3.4.3.2.2 ISMapper analysis

ISMapper was used in typing mode to screen for ISAba1 and compared to the *A. baumannii* BAL062 reference. ISAba1 interruptions were detected in either *lpxC* and *lpxD* in seven isolates (Table 3.2). Four isolates also had interruptions in *mlaA* in addition to the *lpxC* gene (Table 3.2). The gene *mlaA* encodes for a membrane transporter[337], and the interruption of this gene could disrupt the membrane and cause resistance to colistin. Both isolates (45A and 45B) which contained a SNP in their *lpxA* gene, also had ISAba1 insertions within their *lpxC* genes.

Table 3.1: Polymyxin B resistance mutations in each *A. baumannii* isolate in this study (adapted from Lim *et al.*, 2015).

Isolate	Gene	Product	Mutation	Location
1	*pmrB*	Two-component sensor kinase signal peptide	P233S	YP_003730963.1
2	*pmrB*	Two-component sensor kinase	R263H	YP_003730963.1
	mutS	DNA mismatch repair protein	IS*15* insertion	1792809–1792801
3	*lpxA*	UDP-N-acetylglucosamine acyltransferase	ISAba1 insertion	2690056–2690049
4	*lpxC*	N-Acetylglucosamine deacetylase	ISAba1 insertion	3758548–3758539
5	*lpxC*	N-Acetylglucosamine deacetylase	ISAba1 insertion	3758550–3758559
6	*lpxD*	UDP-3-O-(3-hydroxymyristoyl) glucosamine N-acyltransferase	S167F	YP_005515126.1
7	*lpxA*	UDP-N-acetylglucosamine acyltransferase	G56V	YP_003731736.1
8	*lpxC*	N-Acetylglucosamine deacetylase	ISAba1 insertion	Amino acid positions 155–159
9	*lpxA*	UDP-N-acetylglucosamine acyltransferase	Stop codon E84*	YP_003731736.1
10	NA	Putative metal transporter	Codon deletion HH171	YP_005523968.1
	lpxC	N-Acetylglucosamine deacetylase	Frame shift with a 7-bp insertion resulting in incomplete codon insertion (Q252QSS)	YP_047978.1

Table 3.2: Non-conserved IS*Aba1* insertions within each pair of colistin passaged *A. baumannii* isolates.

Isolate	Colistin concentration	Gene	IS*Aba1* location
27A	32 mg/L	-	-
27B	32 mg/L	-	-
32A	64 mg/L	*lpxC*	163013 - 162958
		lpxD	1686561 - 1686610
		baeR	613745 - 613706
32B	64 mg/L	*lpxC*	163013 - 163022
		mlaA	3357969 - 3357991
38A	64 mg/L	-	-
38B	64 mg/L	*lpxC*	163011 - 163020
		baeS	612247 - 612126
35A	128 mg/L	*lpxC*	163013 - 162995
		mlaA	3357966 - 3358025
35B	128 mg/L	*lpxC*	163013 - 162995
		mlaA	3357966 - 3357969
		between BAL062_00181 and *pleD*	206902 - 206911
45A	128 mg/L	*lpxC*	163007 - 163011
		lpxC	163329 - 163338
45B	128 mg/L	*lpxC*	163011 - 163013
		lpxC	163409 - 163260
		mlaA	3358299 - 3358301

3.4.4 Common pathways to IS-mediated polymyxin resistance

Within *A. baumannii*, multiple different mechanisms of IS-mediated resistance to polymyxins were observed. In the polymyxin B passage study, the majority of resistance to polymyxin B was due to SNPs or indels within *pmrB*, *lpxC* or *lpxA*. In some cases, ISAba1 was responsible for this resistance by disrupting *lpxC* or *lpxA*. Comparatively, few SNPs or indel mutations were observed in the colistin passage study, with only two isolates containing SNPs in *lpxA*. The majority of colistin resistance in this dataset were due to ISAba1 interruptions in *lpxC* and *lpxD*. The membrane gene *mlaA* was also interrupted in four isolates, potentially contributing to colistin resistance. In both studies ISAba1 was the primary IS responsible for resistance, and most commonly interrupted the *lpx* genes. The only exception was an insertion of IS*15* within *mutS*, potentially generating a hypermutator phenotype that may be advantageous to adapting to high levels of polymyxins.

3.5 Discussion

Here I have shown how ISMapper can be used to identify mechanisms of antibiotic resistance in bacterial pathogens of human health importance. These mechanisms were not explained by more common analysis approaches aimed at identifying resistance-associated SNPs or horizontally acquired resistance genes. Four distinct mechanisms for IS-mediated antibiotic resistance were identified using ISMapper.

Firstly, section 3.2.2 demonstrates that IS can be effective at restructuring AMR loci, such as the SGI. In the SGI, IS*26* was responsible for rearrangements in the SGI, as well as IS-mediated loss of resistance genes within this region. In section 3.3, IS*1* was shown to be critical for incorporating the AMR transposon Tn*2670*, normally located in the MDR IncHI1 plasmid, into the chromosome. Section 3.4.2 demonstrated the spread of carbapenem resistance in *A. baumannii* through multiple, independent acquisitions of *oxa23* transposons flanked by IS. Further, IS were demonstrated as responsible for conferring resistance by upregulating gene expression within these genomes. The insertion of ISAba1 and the subsequent upregulation of *ampC* in *A. baumannii* resulted in AMR to third generation cephalosporins. Finally, section 3.4.3 demonstrated how IS-mediated gene interruption has played a crucial role in resistance to the drugs of last resort, polymyxins. The interruption of important membrane genes *lpxA*,

lpxC, *lpxD*, and possibly also *mlaA*, conferred resistance to polymyxin B and colistin. In all of these scenarios, ISMapper was instrumental in identifying the distinct IS-mediated causes of antibiotic resistance.

ISMapper will continue to play a significant role in teasing apart the complexity of AMR in bacterial pathogens going forward. Increasingly, bacterial pathogens are being investigated using short reads[191,250,264], and using these reads to perform epidemiological surveillance or screening in a clinical setting[192,338,339]. Short reads are difficult to use for detecting IS. In the future, long read sequencing such as PacBio or Nanopore may be able to overcome this issue (as seen in *S.* Typhi, section 3.3), but currently long read sequencing is expensive to implement[340]. Whilst long read sequencing costs will come down over time, it may never be affordable for use in a clinical setting. In addition, there has already been a large amount of short read data made publicly available through projects such as GenomeTrackr, and the amount of short read data will only increase over time. Given this, ISMapper still has a significant role to play in teasing apart the complexity of antibiotic resistance in bacterial pathogens going forward.

In addition to detecting resistance mechanisms in a public health and epidemiological settings, it is important to continue understand how antibiotic resistance continues to evolve and disseminate. To achieve this, further research is required to investigate different resistance phenotypes that arise via IS-mediated mechanisms. Finding the genetic context of resistance regions, such as transposons or large islands, will be key to understanding how they spread through bacterial populations. Tools such as ISMapper can assist with these lines of inquiry.

Chapter 4

Dynamics of insertion sequences in *Shigella sonnei*

4.1 Introduction

S. sonnei and the three other *Shigella* species are human-adapted lineages of *E. coli*[221]. A recent study performed by Holt *et al.*[250] sequenced 132 *S. sonnei* genomes on the Illumina Genome Analyzer GAII. The isolates in this study came from 27 countries across four continents, spanning the years 1943 - 2008. Using these genomes, the study elucidated the global population structure of *S. sonnei*, showing that it arose in Europe in the mid to late 17th century and has since evolved into four lineages, three of which contain multiple strains from the 1940s to the 2000s, and one lineage comprising of a single genome from France[250]. The three major lineages diverged from their common ancestor in the early to mid 19th centuries, with isolates from lineages I and II mostly being confined to Europe. Of the three major lineages, lineage III, is now the most prevalent, due largely to the global dissemination of a single MDR subclade known as Global III[250,252]. The mean substitution rate across the whole population was estimated to be 6.0 x 10^{-7} substitutions per site per year[250].

Shigella genomes harbour hundreds of pseudogenes and large numbers of IS compared to *E. coli*[341]. IS have been an important part of the evolution of *Shigella* from *E. coli* – they have been responsible for the inactivation of motility genes, and genes that hinder *Shigella*'s ability to cause disease (as discussed in Chapter 1). Each of these gene inactivations have contributed to its patho-adaptation in humans[95].

4.1.1 Aims

This chapter investigates the evolutionary dynamics of IS in the previously published *S. sonnei* genomes[250]. This chapter uses the phylogenetic structure inferred in Holt *et al.*[250] to assess IS dynamics across the global population. From the 132 genomes in Holt *et al.*[250], only 126 of these are included in this chapter. The single lineage IV isolate was excluded from this study, as there was only one representative. Five additional genomes were excluded as they were low-depth, with < 10x coverage, so IS insertion sites could not be detected accurately by ISMapper (as discussed in Chapter 2). The *S. sonnei* 53G reference genome was screened for IS, and 12 different IS elements were identified. All 126 genomes were then interrogated for these 12 IS.

The specific aims of this study were:

i) To determine which IS are present in the global population of *S. sonnei*, and investigate

differences in IS content or dynamics between the three lineages;

ii) To examine the evolutionary history of IS in *S. sonnei*; and

iii) To investigate the contribution of IS to gene inactivation within the global population of *S. sonnei*.

IS insertion sites detected using ISMapper were compared across all genomes to examine burden of IS in individual genomes and lineages (section 4.3.1). Using ancestral state reconstruction, the burden of IS in ancestors of the *S. sonnei* lineages was inferred, and used to examine the rates of gain and loss of IS across the evolutionary history of *S. sonnei* (section 4.3.2). The contribution of IS to gene inactivation was compared to other types of mutations (section 4.3.3), the evidence for the contribution of IS inactivation to balancing and negative selection within *S. sonnei* was assessed (section 4.3.4), and the impact of IS on functional diversification of the three *S. sonnei* lineages was investigated (section 4.3.5).

4.2 Methods

4.2.1 Selection of IS and creation of IS-free reference genome

The lineage II *S. sonnei* 53G genome (accession NC_016822) was selected as the reference genome for this study, as this genome was sequenced at the Sanger Institution using capillary sequencing at a higher accuracy level than the only other complete *S. sonnei* genome, strain Sso46. To allow for more accurate detection of precise IS insertion sites within genes, an IS-free version of this reference was created as follows.

ISSaga[342] was used to detect all IS within the complete *S. sonnei* 53G reference genome. ISSaga screens the genome for homologs of know IS present in the ISFinder database. IS present in more than one copy, with at least 80% nucleotide identity to an IS in the ISFinder database, were selected for downstream screening. These IS were annotated using the ISSaga output, and manually inspected to ensure that the IS sequence was complete, and any target site duplications surrounding the IS were included. Each IS and its target site duplication (if present) was removed from the reference genome sequence to aid accurate detection of IS-mediated gene interruption. Annotations of features (including CDS and gene features) in

the complete reference genome were transferred from the complete 53G reference genome to the IS-free 53G reference genome using RATT[343], with the strain transfer parameter.

Functional assignments for *Shigella* genes were extracted using RAST to annotate protein-coding sequences using the SEED database, a curated database of protein families, called FIGfams, which organises genes into functional categories, subcategories, and subsystems[344]. Initially, genes were compared to the curated database of *E. coli* K-12 genes, EcoCyc, in an attempt to assign functional categories to inactivated genes by extracting Gene Ontology (GO) terms. The majority of the GO terms provided were relatively uninformative high-level terms (eg: 'membrane'). Therefore, to obtain high quality functional assignments for *Shigella* genes, RAST was used instead, which relies on the SEED database. Existing gene annotations in the IS-free 53G reference genome (doi: 10.4225/49/589c2ef7128ac) was annotated again using RAST[345], preserving current gene calls, to obtain FIGfam numbers for each gene. Genes without a FIGfam number could not be assigned to a functional category.

4.2.2 Detection of IS, nonsense SNPs and indels

The *S. sonnei* data consisted of 126 genomes from the Holt *et al.*[250] study, sequenced on the Illumina Genome Analyzer GAII, generating paired end reads. Sequenced genomes had an average read length of 59 bp. The average read depth was 83x (range 68x to 91x). All 126 genomes were screened for the 12 IS identified by ISSaga (section 4.2.1, IS*1*, IS*2*, IS*4*, IS*21*, IS*600*, IS*609*, IS*630*, IS*911*, ISEc20, ISSso1, ISSso4 and ISSso6) using ISMapper in typing mode, against the IS-free 53G reference genome, using default settings (minimum depth of 6x for detection of an IS insertion site).

All 126 genomes were mapped to the IS-free 53G reference genome using RedDog pipeline v01.9b (see section 3.2.3.2 for an explanation of this pipeline) to detect nonsense SNPs (SNPs which create a premature stop codon) and indels. Indel positions were extracted from the VCF files output by RedDog, and these indel positions were compared to the annotations in the IS-free 53G reference genome to determine which indels were within genes. Only indel positions causing frameshifts within genes were kept for downstream gene inactivation analysis.

To identify nonsense SNP mutations, the SNP consequences file from RedDog was used, which annotates each SNP with its coding effect (i.e. whether it is intergenic or genic; and for genic SNPs what the effect on the encoded protein is), based on the annotation of the coding

features in the reference genome (in this case, the IS-free 53G reference genome). From this SNP consequences file, only SNPs generating nonsense mutations were extracted and kept for downstream gene inactivation analysis.

4.2.3 Assembly of all *S. sonnei* genomes

All *S. sonnei* genomes were assembled using SPAdes v3.6.2[304] with kmer sizes 21, 33 and 55. The contigs for each genome were ordered against the complete *S. sonnei* 53G reference genome using Abacas v1.3.1[307] and annotated with Prokka v1.11[308]. To determine genome size, contigs < 200 bp were discarded, as these likely represent small errors in the genome assembly. For each genome, all contigs were concatenated and genome size identified using the infoseq command in the EMBOSS package[346].

4.2.4 Phylogenetic trees used in this study

The phylogeny from Holt *et. al.*[250] was generated by inputting the SNP alignment generated in their study into BEAST v1.6[299]. They used a GTR+Γ substitution model, with a lognormal relaxed clock and a coalescent population size[250]. Ten chains of 100 million iterations were combined, burn-in removed, and summarised into a MCC tree[250]. The six genomes exluded in this chapter were removed from this phylogeny, and the resulting phylogeny was used for the ancestral state reconstruction of IS insertion sites, nonsense SNPs, and indels, as described in section 4.2.5.

A set of random 100 BEAST trees were generated for this study to account for uncertainty in tree topology when estimating rates of IS gain and loss. The SNP alignment output by RedDog was used as input for BEAST v1.8.3[299] with the same model (GTR + Γ substitution model and a lognormal relaxed clock) used by Holt *et. al.*[250], with five independent chains of fifty million iterations each. Each chain had ESS values of > 200 for all parameters, with good sampling of the likelihood surface. All five chains were combined and subsampled using LogCombiner v1.8.3[299], by removing 500,000 trees as burn-in and resampling every 100,000th tree. From the remaining 2475 trees, 100 trees were selected at random, to represent the population of trees, for further analysis in section 4.2.5.

Finally, additional lineage specific trees were generated to determine the mutation rate for

each lineage. Individual lineage SNP alignments were extracted from the SNP alignment output by RedDog. To assess the strength of the signal for mutation rate in each lineage, ten replicate alignments with randomised tip dates were created for each lineage. All alignments were analysed separately using BEAST v1.8.3[299] with a GTR+Γ substitution model, a relaxed clock model and a coalescent tree prior. Each chain was run for fifty million iterations, with ESS values > 200. Ten percent burn-in was removed from each run before extracting the mutation rate using Tracer v1.6[300].

4.2.5 Ancestral reconstruction of IS insertion sites, nonsense SNPs, and indels

The presence or absence of each IS site, nonsense SNP, and indel was determined on each internal node of the phylogeny produced by Holt *et. al.*[250] (described in section 4.2.4), using maximum parsimony ancestral state reconstruction, implemented in the ancestral.pars function in the R package phangorn v2.1.1[347]. For each IS site, the number of events inferred across the tree (either gain or loss) was calculated as follows. For nodes where the IS insertion was inferred to be absent, but inferred as present on its parent node, a loss event was recorded. For nodes where the IS insertion was inferred to be present, but inferred as absent on its parent node, a gain event was recorded. If there was no change in the inferred IS state between the current node and the parent node, then no event was recorded. These results were collated to determine the total number of gain and loss events occurring on each branch (excluding the deep branches leading to lineage I and the mrca of lineages II and III, due to uncertainty in the ancestral state reconstruction at these nodes), across all IS insertions.

To account for uncertainty in the tree topology, this ancestral state reconstruction was repeated on the set of 100 random BEAST trees generated in section 4.2.4.

To summarise the results of IS gain and loss over time, each branch in each tree was binned into the decades it spanned (with some long branches spanning multiple decades). The number of events on each branch, in each decade interval, was assumed to occur at a constant rate over time, and so the events were assumed to be evenly distributed across the entire length of the branch. To estimate the number of events in each decade interval (e_d), the total number of events on each branch that spans that interval (N) was weighted by the proportion of the branch length that overlaps the interval (y_d).

To adjust the number of gains and losses by decade, the following formula was used:

$$e_d = \frac{N}{y_t} * y_d$$

where

e_d, the adjusted number of events occurring on the branch in a particular decade;

N, the total number of gains or losses inferred on the branch;

y_t, the total length of the branch in years;

y_d, the number of years the branch spans within that decade.

To smooth the resulting data, the five moving mean method was used - the mean of the median number of gains or losses in a particular decade (i), plus the median gains or losses in the two neighbouring decades on either side ($i - 2$ to $i + 2$), was taken.

4.2.6 Modelling of changes in IS copy number over time

Initial modelling of IS copy number was performed using a linear regression, comparing the overall IS copy number in extant genomes with the year the genome was isolated. The results of the linear regression were used to inform a logistic growth model of IS copy number growth over time, using inferred dates and copy numbers for internal nodes of the tree, in addition to extant genomes. In this model, the rate of IS gain over time was assumed to fit a logistic function, where initial gain of IS copies was large, before decreasing in rate and reaching a saturation point. A Bayesian MCMC approach was designed to fit this model to the data.

A sub-tree for each lineage was extracted from the overall *S. sonnei* phylogeny. The root of each sub-tree was set to time zero. Time of each node relative to the root was calculated for each node and tip in each of the sub-trees were calculated using the allnode.times function in the R package NELSI (https://github.com/sebastianduchene/NELSI). To fit the logistic function, two assumptions were made:

i) that lineage I, with the largest slope from the linear regression, was still undergoing rapid IS expansion and would sit on the steeper section of the curve; and

ii) that lineages II and III, with weaker linear regression slopes, were approaching IS saturation and would sit ahead of lineage I on the curve.

To model these assumptions, lineage I time points were left as is (with the mrca of lineage I set to time zero), but lineage II and III points were offset compared to lineage I, by adding the maximum lineage I time value (172 years) to all lineage II and III points. IS copy number at each node was calculated by summing the inferred ancestral states at each node (see 4.2.5).

The logistic function was defined as:

$$S = \frac{L}{1 + e^{-k(x+x_0)}}$$

The parameters of the model were:

L, the maximum value of the logistic curve;

k, the steepness of the curve;

x_0, the inflection point on the curve.

Additional parameters in the MCMC likelihood function included:

er, an error term to capture the uncertainty in the model;

x_1, the position of lineage I points on the *x*-axis;

x_2, the position of lineage II points on the *x*-axis;

x_3, the position of lineage III points on the *x*-axis.

Priors with normal distributions were placed on all parameters: L, mean = 400, standard deviation (sd) = 10; k and x_0, mean = 0, sd = 10; er, mean = 100, sd = 100, x_1, x_2 and x_3, mean = 0, sd = 0.1. Tight priors were placed on the parameters identifying the *x* positions of each point to control for non-identifiability.

Ten independent runs of the chain were performed, with ten million iterations each. Burn-in (100,000 steps) for each run was removed, and all ten runs were combined and subsampled to produce the final estimates for each parameter. Ten independent replicates of the chain were performed using no data to estimate the distributions of the prior and ensure that the prior functions were not overly informative, and that the data was driving parameter estimates.

4.2.7 Identification of genes under balancing or negative selection

To identify genes under negative selection, two different methods were used to differentiate between genes under negative selection, and genes undergoing gene decay.

Firstly, to identify genes under negative selection, rather than a lack of purifying selection, the number of IS insertion sites within each gene was modelled using a Poisson distribution, with mean rate lambda (λ) equalling the intergenic insertion rate, defined as the number of intergenic insertions (i_i) divided by the number of intergenic bases (i_b) in the IS-free reference genome. The intergenic rate was was used as an estimate of overall insertion rate, based on the assumption that most insertions outside of coding genes are likely to be neutral. However, it is likely that some intergenic insertions may have functional effects that are subject to purifying selection, so this can be thought of as testing for genic insertions that are subject to greater purifying selection than that on intergenic sites.

For each gene, the number of insertions observed in the gene was compared to the Poisson cumulative distribution function to estimate the probability that the gene had a higher number of IS insertion sites than expected by chance, given the size of the gene. All p values were then corrected using the FDR method, and genes with an adjusted p value of > 0.05 were considered to be under negative selection.

Secondly, to determine which genes were under negative selection but not undergoing gene decay, an insertion index (ii) for each gene was calculated and corrected for multiple gene inactivation events. The insertion index was calculated by dividing the number of IS insertion sites identified within the gene, across all isolates, by the length of the gene in base pairs.

$$ii = \frac{\sum sites}{length}$$

The rate of IS insertion within a gene varies depending on gene length. A high insertion rate alone does not delineate between very short genes, with only one or a few IS insertions, and genes with high numbers of IS insertions within them. For example, a 200 bp gene with 1 IS insertion would have an insertion rate of 0.005, and a 1000 bp gene with 5 IS insertions would have the same insertion rate of 0.005. Additionally, genes with a large number of IS insertions could be generated under two very different evolutionary scenarios:

i) each IS insertion is an independent inactivation event, indicating that this gene is under negative selection; or

ii) there has been an initial gene inactivation event, followed by more insertions within the gene, indicating that this gene is undergoing gene decay.

To correct the insertion index and identify genes accumulating inactivation events within the same genome, rather than having a high number of independent inactivation events, an inactivation frequency was calculated. Inactivation events were defined as inactivating IS insertions, inactivating indels, and nonsense SNPs. For each gene, the number of genomes with more than one inactivation event (either IS insertion or mutational inactivation) was counted (g_m). This was divided by the total number of genomes (n=126) and multiplied by the insertion index to generate an inactivation frequency (i_f).

$$i_f = \frac{g_m}{126} * ii$$

To determine the amount of degradation a gene was undergoing, a degradation index was calculated. For each gene, the number of genomes containing an inactivation event were counted, so some genomes were counted multiple times if they contained multiple inactivations within the same gene (g_i). The total number of genomes with more than one inactivation event (g_m) was subtracted from the total number of genomes containing an inactivation (g_i). This result was then divided by g_m, the number of genomes with multiple inactivation events. Genes with a degradation index of 1 were considered to be highly degraded, and genes with a degradation index of 0 were considered to be un-degraded.

$$di = \frac{g_i - g_m}{g_m}$$

To determine which genes have a high number of inactivations due to multiple independent inactivation events, and which genes are accumulating inactivations, the degradation index and insertion frequency were combined. The resulting measure, (s), summarises this and is calculated as follows:

$$s = (1 - di) * (\frac{g_m}{126} * ii)$$

Genes with a value of $s > 0.001$ were considered to be un-degraded, and undergoing either balancing or negative selection.

4.2.8 Identifying RAST categories enriched for inactivated genes

To explore the relationship between gene inactivation and functional categories, two-way contingency tables (example table below) were constructed for each RAST category identified in section 4.2.1.

Table 4.1: Example contigency table for odds ratio calculation.

	inactive gene	**active gene**
in RAST category	A	B
not in RAST category	C	D

Fisher's exact test (function fisher.test in R) was used to calculate odds ratios and confidence intervals (two-sided tests in all cases). P values were calculated using the function phyper, and all p values were adjusted for multiple testing using the FDR method.

4.2.9 Detection of colicin genes

All 126 genomes were assembled as previously described (section 4.2.3). To generate a pangenome, MUMmer v3.23[348] was used to compare all assemblies to each other, adding novel sequences > 100 bp to the pangenome. The pangenome was then annotated using Prokka v1.11[308]. A tBLASTn search using colicin E protein sequences obtained from NCBI was performed on the pangenome to identify any colicin E type gene (hit defined as >70% identity and >60% coverage), however, only colicin E1 and colicin E3 toxin and immunity genes were detected. All genomes were mapped to the pangenome using RedDog v1.09b to determine presence/absence of all genes. The presence/absence table output by RedDog defines gene presence as 95% of the reference gene covered with a depth of > 5 reads. Using this output, the presence of colicin E1/E3 toxin and immunity genes was determined for each genome.

4.3 Results

4.3.1 IS burden in *S. sonnei* lineages

Twelve distinct IS detected in the complete *S. sonnei* 53G reference genome were identified, and their insertion sites were determined in each of the 126 genomes using ISMapper. A total of 1,227 insertion sites were detected across all isolates (Table 4.2). IS load varied markedly by lineage: lineage I genomes had the fewest IS per genome (median 243), while lineages II and III had significantly higher IS counts than lineage I (median 299 and 294 respectively, $p=1.97x10^{-10}$) (Figure 4.1d-f). Each of the 12 IS were identified in at least one copy in every genome, and each of the 12 IS were found in similar proportions in all genomes (Figure 4.1a), however the insertion sites differed sufficiently between lineages to allow clustering of isolates into lineages using the first two principal components extracted from IS insertion site profiles (Figure 4.1c, Figure 4.2).

IS*1* contributed the greatest to overall IS burden in *S. sonnei* (median 48% of IS insertions per genome, see Figure 4.1a). IS*2*, IS*600* (IS*3* family) and IS*4* (IS*4* family) were also abundant (median 10%, 14%, 10%, respectively). There were relatively few copies of IS*911* (IS*3* family); ISSso1, ISSso6 and ISEc20 (IS*110* family); IS*21* and ISSso4 (IS*21* family); IS*630* (IS*630* family); and IS*609* (IS*200*/IS*605* family) (Figure 4.1a). Only three insertion sites were detected for IS*609*, and all three were found in all lineages but lacking from some individual strains in each lineage (Figure 4.2). Therefore, all three IS*609* sites are hypothesised to have been present in the most recent common ancestor (mrca) of *S. sonnei* and have subsequently undergone occasional loss but not transposition to new locations. All other IS show evidence of ongoing transposition activity: assuming strain-specific insertions represent recent transposition events arising on terminal branches, IS*1* displayed the most activity, followed by IS*2*, IS*600*, IS*911* and IS*630* (Figure 4.1b).

The global genome collection includes isolates sampled over several decades (1943 - 2008), and linear modeling of IS copy number on year of isolation indicated a positive linear relationship for all lineages (Figure 4.3a). However, this relationship was only significant for lineage II (Figure 4.3a).

Taken together, increasing copy numbers over time, and high numbers of strain-specific insertion sites, indicate that there is an ongoing accumulation of IS through transposition.

Table 4.2: IS detected in 126 *S. sonnei* genomes using ISMapper analysis.

IS	IS family	# sites across 126 genomes	mean proportion per genome	# strain-specific insertions
IS*1*	IS*1*	495	0.49	220
IS*2*	IS*3*	199	0.1	128
IS*600*	IS*3*	188	0.14	74
IS*630*	IS*630*	106	0.05	54
IS*911*	IS*3*	76	0.016	62
IS*4*	IS*4*	71	0.1	21
ISSso4	IS*21*	29	0.013	19
IS*21*	IS*21*	22	0.034	5
ISSso6	IS*110*	17	0.02	4
ISEc20	IS*110*	13	0.03	3
ISSso1	IS*110*	8	0.01	3
IS*609*	IS*200*/IS*605*	3	0.01	0

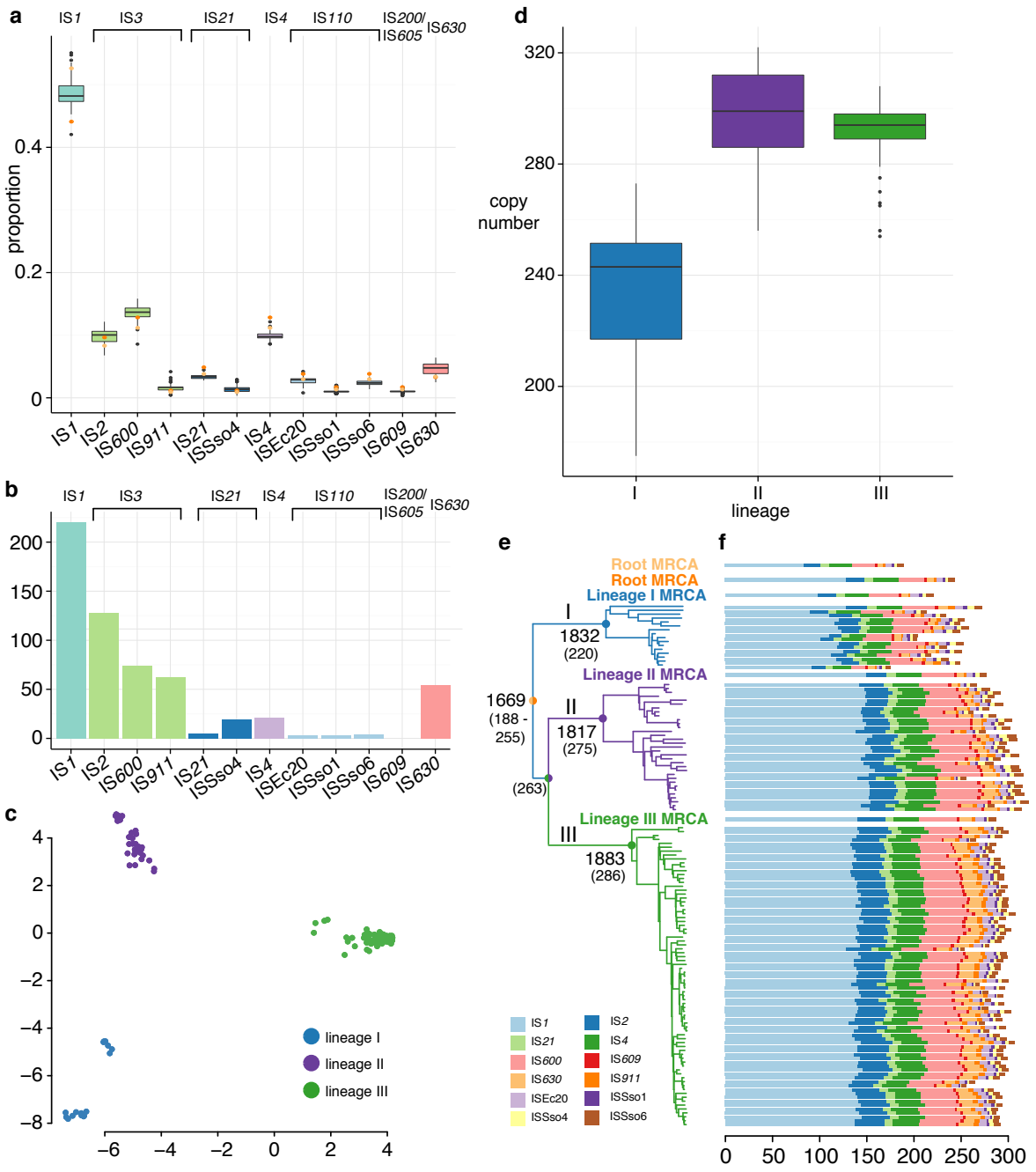

Figure 4.1: Burden of IS in *S. sonnei*. a, Box plots showing overall proportion of each IS within *S. sonnei* genomes. IS family is indicated across top. Orange dots indicate the proportion of that IS within the most recent common ancestor (mrca) of *S. sonnei*. **b,** Number of strain-specific insertions per IS, with bars coloured by IS family, and IS family printed across top. **c,** PCA plot of IS profiles (as shown in Figure 4.2) for each genome. Points are coloured by lineage as per legend. **d,** Box plots of total IS copy number within each lineage. **e,** Phylogeny of *S. sonnei* with lineages labelled and coloured. Marked nodes indicate year of most recent mrca with number of IS copies indicated underneath in brackets. **f,** Stacked bar plots of IS copy number in each genome, and mrcas, coloured by IS as per legend.

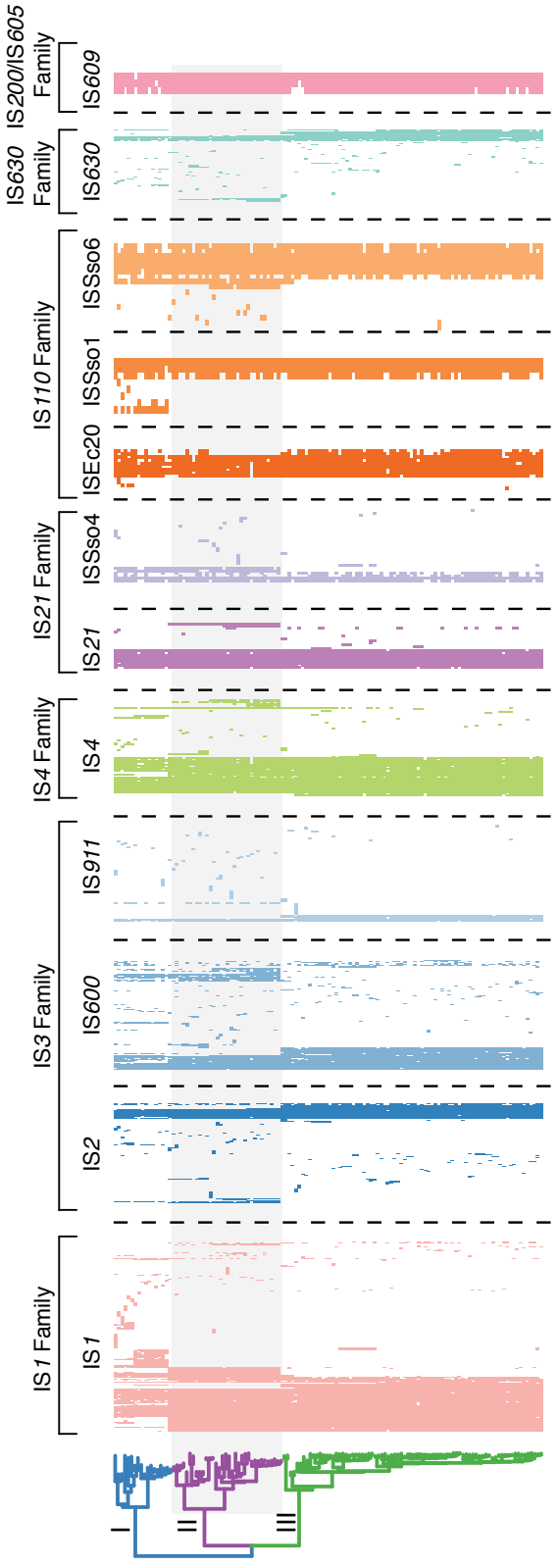

Figure 4.2: Presence of IS insertion sites in individual *S. sonnei* genomes. Each row represents a genome, corresponding to the tip in the BEAST tree (left). Each column respresents a specific IS insertion site; coloured blocks indiciate presence of the IS insertion in each genome. Columns are clustered by IS, not ordered by position within the genome.

4.3.2 Evolutionary history of IS in *S. sonnei*

In order to investigate the historical patterns of IS in *S. sonnei*, maximum parsimony ancestral state reconstruction was used to infer the presence of each IS insertion at internal nodes of the dated phylogeny. Transitions between presence and absence at linked nodes were assumed to represent IS gain and loss events, and total IS burden was calculated at each node (see Methods 4.2.5). For each IS, the total number of inferred gains and losses were similar in magnitude, with a slight excess of gains for the most common IS, suggesting that these IS are still accumulating (Figure 4.4a). The total number of gains across the tree was also positively correlated with the number of gains and strain-specific insertions for each IS (correlation coefficient=0.88, R^2=0.76, p=0.0001, Figure 4.4b), indicating that both measure of activity are in agreement.

The ancestral state reconstruction data were used to estimate IS copy number at internal nodes. Based on this analysis, the mrca of lineage I, which existed circa 1832, was inferred to carry 220 IS insertions (Figure 4.1e). In contrast, the mrcas of lineages II (circa 1817) and III (circa 1883) were inferred to carry 275 and 286 IS respectively, higher than the copy number of contemporary isolates of lineage I (median 243; see Figure 4.1e, Figure 4.3). The mrca of lineages II and III was inferred to carry 263 IS (Figure 4.1e). Figure 4.3b suggests a fairly constant accumulation of IS over time within each lineage, but with the lineage I mrca starting from a much lower IS count.

It was not possible to reconstruct the number of IS in the *S. sonnei* ancestor (cira 1669), due to the uncertainty in the reconstruction of IS insertions that were present in the lineage II and III mrca but absent in the lineage I mrca (55 sites, Table 4.3), or present in the lineage I mrca and absent from the lineages II and III mrca (12 sites, Table 4.3). Most of the sites that differed between these two nodes belonged to IS*1*, and this was higher than the overall proportion of IS*1* found in *S. sonnei*. For these sites, it was not possible to determine if they were present at the root and lost in one branch, or absent at the root and gained in the other branch. This observation led to two possible hypotheses regarding the number of IS in the *S. sonnei* ancestor:

i) that the IS burden at the root of the tree was similar to that of the lineage I ancestor, and the ancestor of lineages II and III rapidly gained additional IS copies over a short time period, after diverging from the lineage I ancestor (light orange dashed lines, Figure 4.3b); or

ii) that the IS burden in the *S. sonnei* ancestor was similar to the burden found in the ancestor of lineages II and III, with many IS insertions lost in the lineage I ancestor (orange dashed

lines, Figure 4.3b).

Under hypothesis i), the mrca of *S. sonnei* would have contained 208 IS, as these were the IS present in both the mrca of lineage I and the mrca of lineages II and III (light orange dashed line, Figure 4.3b). Subsequently, the mrca of lineages II and III would have gained 55 IS (45 IS*1*) over a 35 year period, while the mrca of lineage I would have gained only 12 IS (8 IS*1*) over 164 year period. In contrast, under hypothesis ii), the mrca of *S. sonnei* would have contained 255 IS, as these were the IS insertion sites present in the mrca of both lineage I and lineages II and III, as well as those IS insertion sites found only in the lineage I mrca but not the lineage II and III mrca, and vice versa (dark orange dashed line, Figure 4.3b). Subsequently, the mrca of lineage I would have lost 35 IS over a 164 year period, while the mrca of lineages II and III would have gained 8 IS over a 35 year period. The breakdown of IS present in the mrca of lineages II and III, but absent from lineage I, showed that 80% of these sites were IS*1*, almost double the overall proportion of IS*1* sites within extant *S. sonnei* genomes (all genomes, 49%; lineage I genomes, 49%; lineage II genomes, 50%; lineage III genomes, 48%).

This study has so far only explored the rate of IS accumulation by comparing the IS burden in extant genomes with their isolation date. This linear regression does not delineate between gain and loss events, and only investigates total IS burden. A more robust approach was required to separate the contribution of IS gain and loss to overall IS burden over time. To explore this, the number of gain and loss events inferred on each branch of the dated phylogeny was estimated, and the relationship between these event counts and the amount of evolutionary time represented by each branch was investigated (see Methods 4.2.5). This analysis was performed separately for each lineage, as they vary in their molecular clock signal and inferred substitution rates (Figure 4.5): Lineage I, 3.35×10^{-7} substitutions site^{-1} year^{-1}, moderate signal; Lineage II, 3.45×10^{-7}, strong signal; Lineage III, 6.45×10^{-7}, strong signal. The results are shown in Figure 4.6, which includes the median and range of estimates obtained from a random sample of 100 trees (see Methods 4.2.5). Over the last ~120 years, the inferred rate of IS gain per decade in lineage I was ~3 per decade, higher than the gain rate inferred in lineages II and III (steady at ~1.5 per decade in lineage III, and increasing from ~1 to ~2 per decade in lineage II; Figure 4.6).

It has previously been suggested that genomes undergoing IS expansion associated with adaptation to a new niche follow a pattern of rapid IS accumulation (as many new IS insertions have a neutral or positive effect by inactivating genes whose activity either offers no advantage

or is disadvantageous in the new niche), which then slows as the genome reaches a saturation point (at which point all intact genes are required in the new niche and further IS activity is mostly disadvantageous). This can be modelled as a logistic growth of IS copy number.

Regardless of why the lineage I ancestor has a much lower IS burden than the lineage II and III ancestors, these lineage differences provide a unique opportunity to compare IS dynamics across the three lineages that appear to be at different stages in their IS trajectory (i.e. at different points on the same logistic copy-number growth curve), with lineage I much further behind than the other two lineages. In order to explore how well a logistic growth model describes the data, all observed (extant) and inferred (ancestral node) IS copy number counts were fitted to a logistic function (see Methods 4.2.6). Since the most recent extant genomes of lineage I have a similar copy number to the mrcas of lineages II and III, the lineage I points were offset from the extant lineage II and III genomes, placing them on the steeper section of the curve (while allowing them to vary during Bayesian MCMC model fitting) (data points, Figure 4.7). The resulting model showed quite good fit for all three lineages (Figures 4.7, 4.8 and 4.9). Each parameter in the model converged, as shown by the trace plots in Figure 4.9. Posterior distributions estimated by the model differed from the initial set prior distributions, confirming that the results are not driven by overly informative priors (Figure 4.8). The model estimated an IS saturation point of approximately 398 IS copies (interquartile range, 364 to 421). This suggests that while IS expansion rates have slowed in lineages II and III, they can be expected to continue to accumulate IS for ~1000 years based on current trends.

Table 4.3: Counts and proportions of IS insertion sites present in either the mrca of lineages II and III, or the mrca of lineage I, and IS insertion sites present in both mrca.

IS	IS Family	proportion lineage II/III mrca	proportion lineage I mrca	proportion in both lineage I and lineage II/III mrca
IS*1*	IS*1*	0.8 (45)	0.67 (8)	0.43 (90)
IS*2*	IS*3*	0.04 (2)	0 (0)	0.09 (12)
IS*600*	IS*3*	0.05 (3)	0.16 (2)	0.14 (30)
IS*630*	IS*630*	0.04 (2)	0 (0)	0.03 (7)
IS*911*	IS*3*	0 (0)	0 (0)	0.01 (2)
IS*4*	IS*4*	0.05 (3)	0.08 (1)	0.12 (24)
ISSso4	IS*21*	0 (0)	0 (0)	0.02 (4)
IS*21*	IS*21*	0 (0)	0 (0)	0.04 (9)
ISSso6	IS*110*	0 (0)	0 (0)	0.04 (8)
ISEc20	IS*110*	0 (0)	0 (0)	0.04 (9)
ISSso1	IS*110*	0 (0)	0.08 (1)	0.01 (3)
IS*609*	IS*200*/IS*605*	0 (0)	0.01 (3)	0.01 (3)

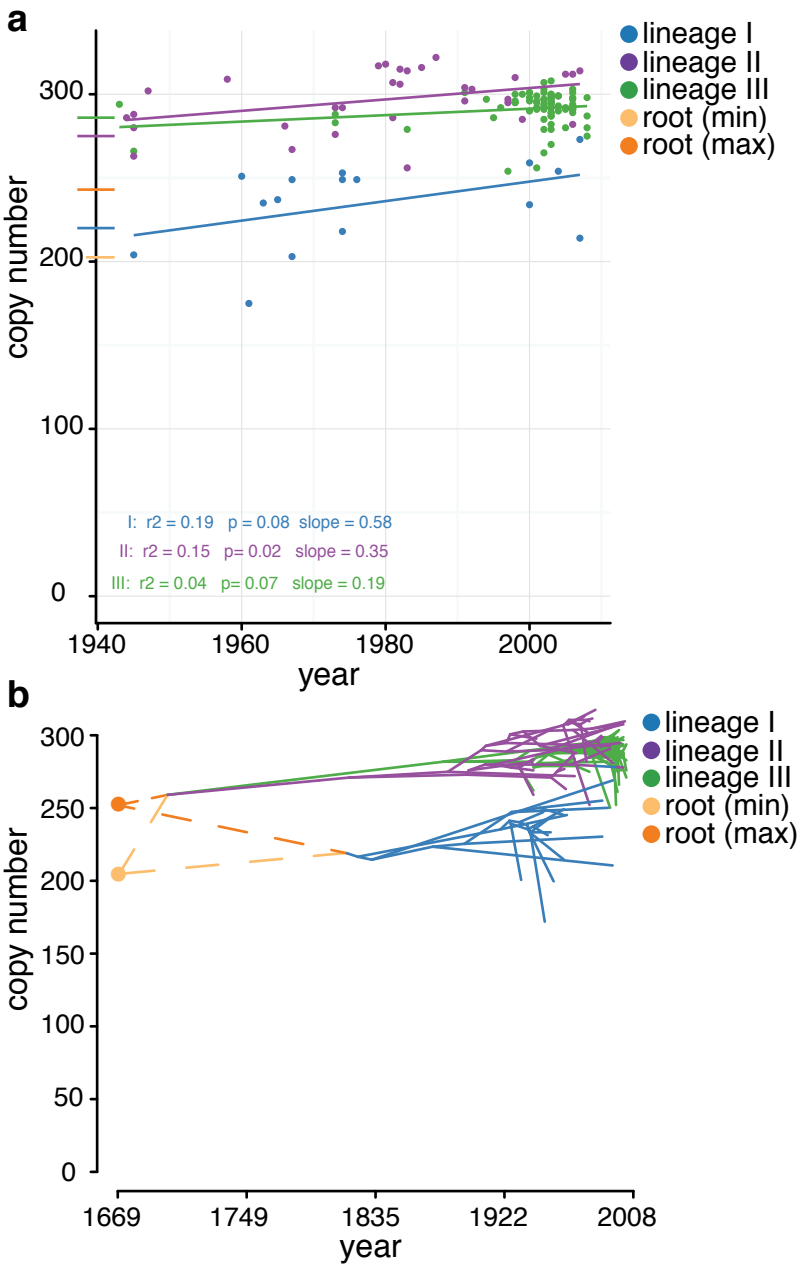

Figure 4.3: Dynamics of IS with *S. sonnei*. **a,** Scatterplot of IS copy number in all extant genomes, with points coloured by lineage. Thick horizontal lines on y axis indicate total IS copy number in each mrca, as per legend. R^2, p and slope values for each regression model indicated in coloured text above the x axis. **b,** Phenogram of tree showing total IS copy number inferred at each node on the tree, except for the mrca of *S. soneni*. Branches coloured by lineage as per legend. Dashed lines indicate the two possible reconstructions from the lower and upper bounded root mrca copy numbers, with lines coloured as per legend.

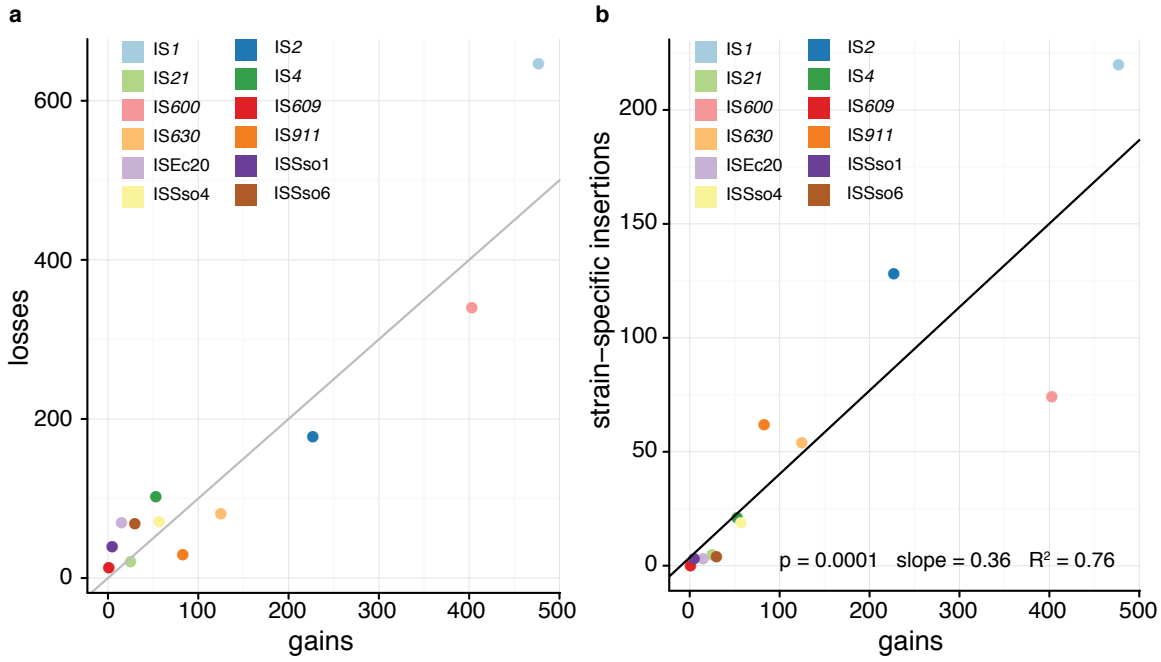

Figure 4.4: Relationship between gain and loss events for each IS inferred across the *S. sonnei* phylogeny, and the relationship between gain events and strain-specific insertions, for each IS. a, Scatterplot showing the relationship between gain and loss events, for each IS, inferred across the phylogeny, using maximum parsimony ancestral state reconstruction (see Methods 4.2.5). Estimates of gain and loss events do not include events at the mrca of *S. sonnei*, the mrca of lineage I, or the mrca of lineages II and III. Dots are coloured by IS as per legend. Grey line shows x=y. **b,** Scatterplot showing total number of gain events against the number strain-specific insertions, for each IS. Dots are coloured by IS as per legend. Black line shows the linear relationship between the number of gain events and strain-specific insertions.

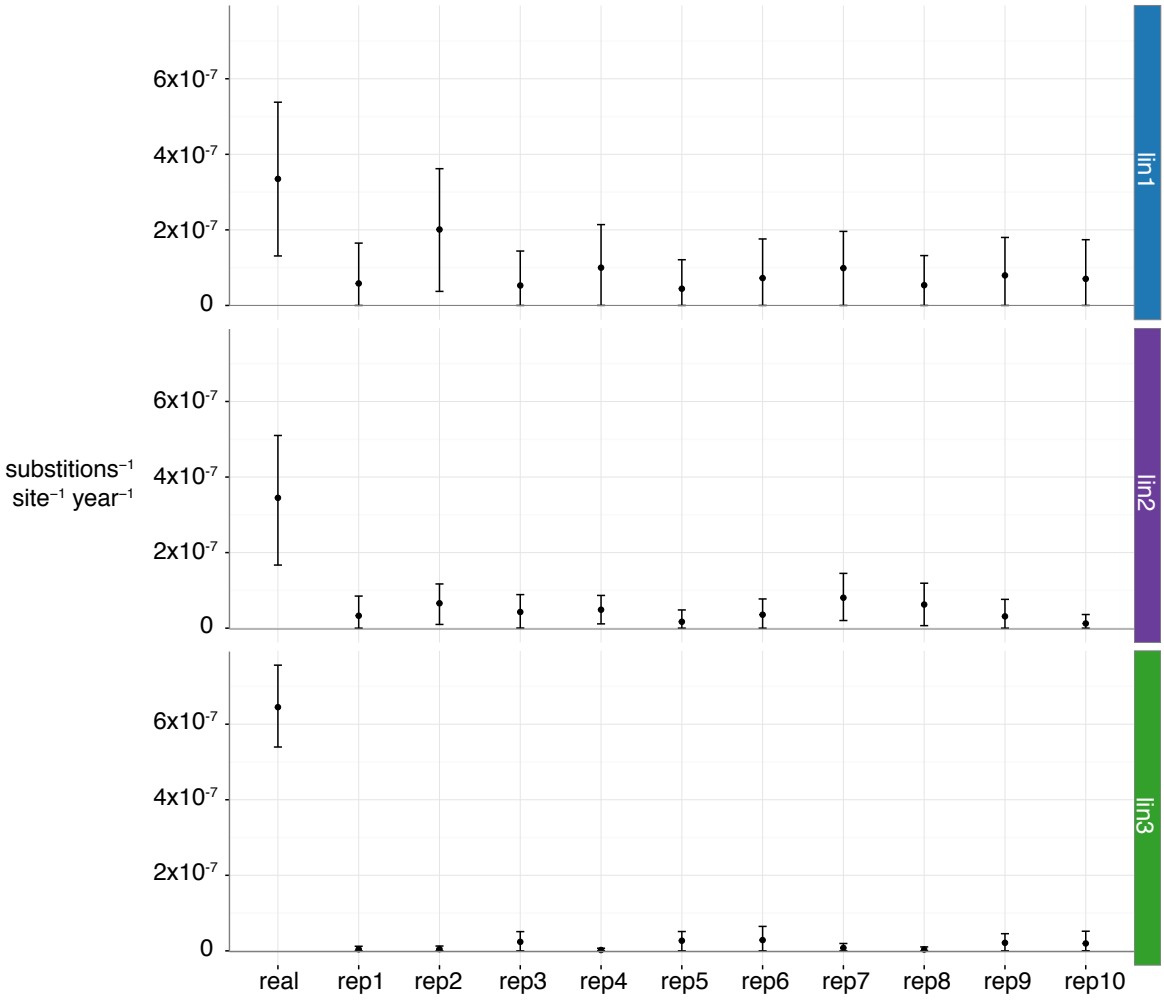

Figure 4.5: Mutation rates in each lineage of *S. sonnei*. Medians (black dots) and 95% HPD intervals (black lines) for mutation rates in each lineage as estimated by BEAST (lineage I, blue; lineage II, purple; lineage III, green). First column in each plot shows the mutation rate estimated using the real isolation dates, with the remaining ten columns showing mutation rates estimated with randomised dates.

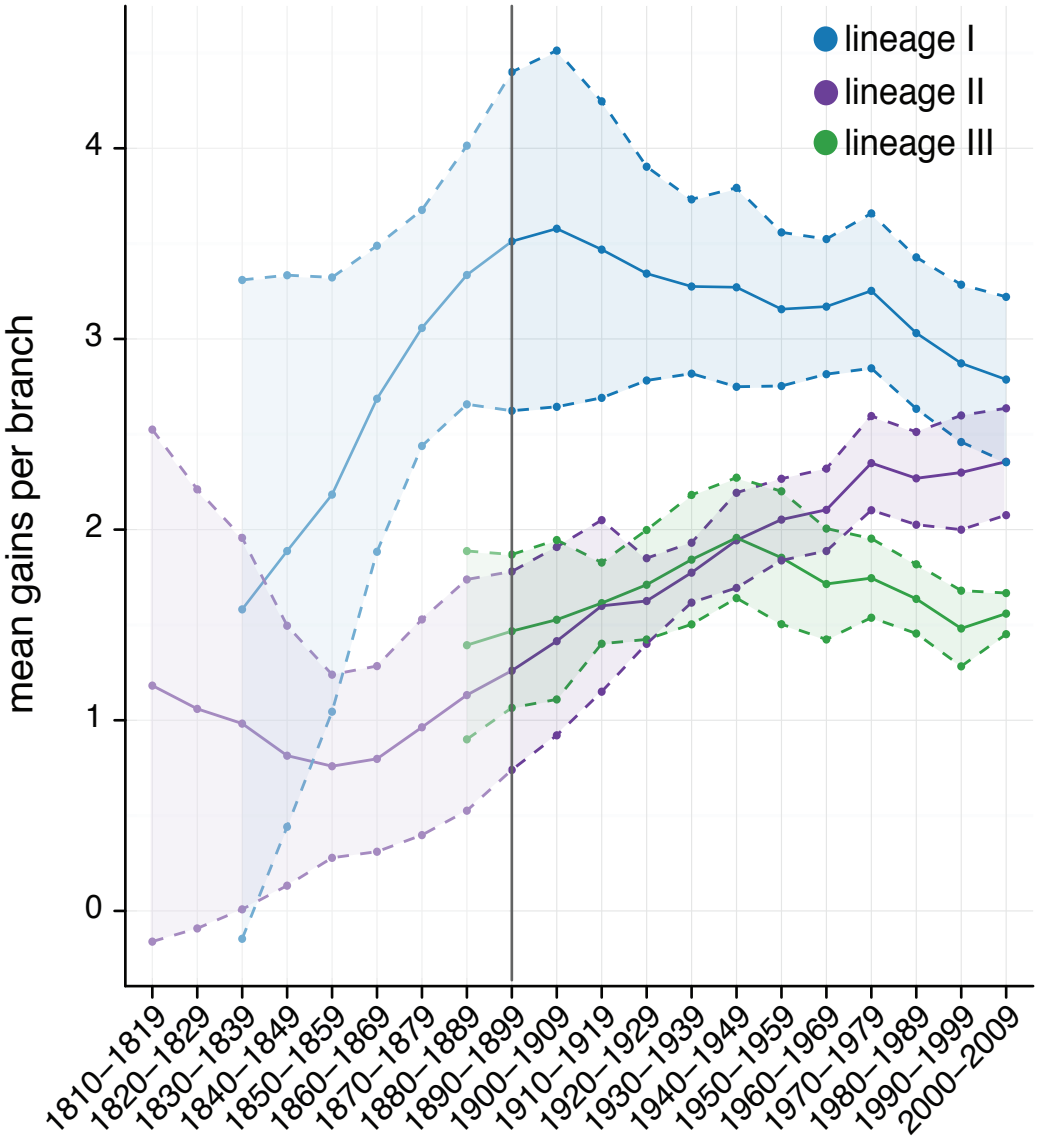

Figure 4.6: Rates of gain in each *S. sonnei* lineage. Mean number number of IS gain events per branch, with branches binned by decade, across 100 random trees. Points and lines are coloured by lineage as per legend. Dashed lines indicate upper and lower interquartile range. Light shaded region (prior to 1900) indicates less confident estimates of gain rate.

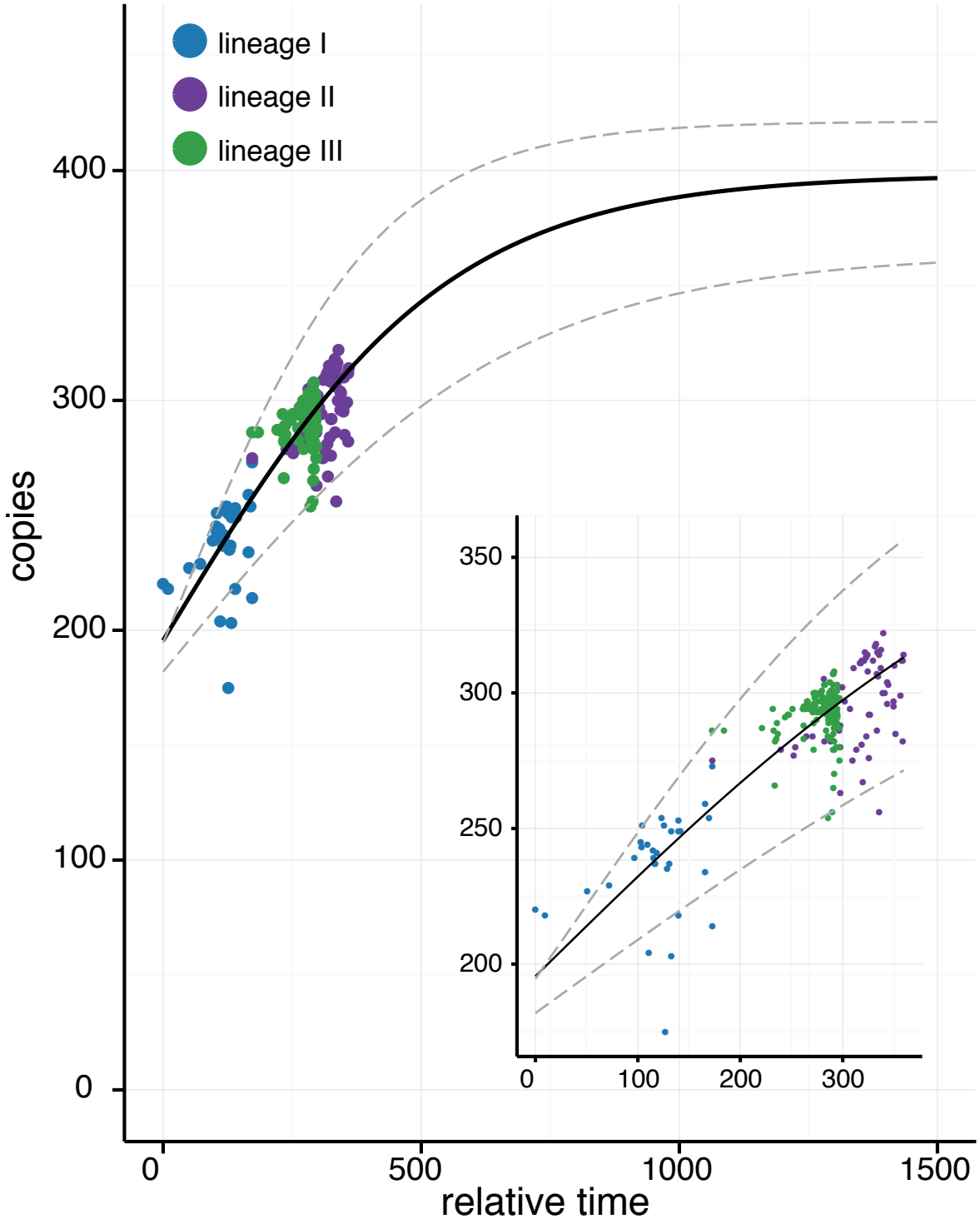

Figure 4.7: Modeling of IS saturation point in *S. sonnei* genomes. Points show the IS copy number of each node and tip in the *S. sonnei* phylogeny against relative time, with points coloured by lineage as per legend. Lineage II and III points are offset in time relative to lineage I. Black line is the logistic curve using parameters estimated by the modeling, with grey dotted lines showing the 95% credible interval of the curve. Insert shows enlarged section of scatterplot, focusing on the lineage points.

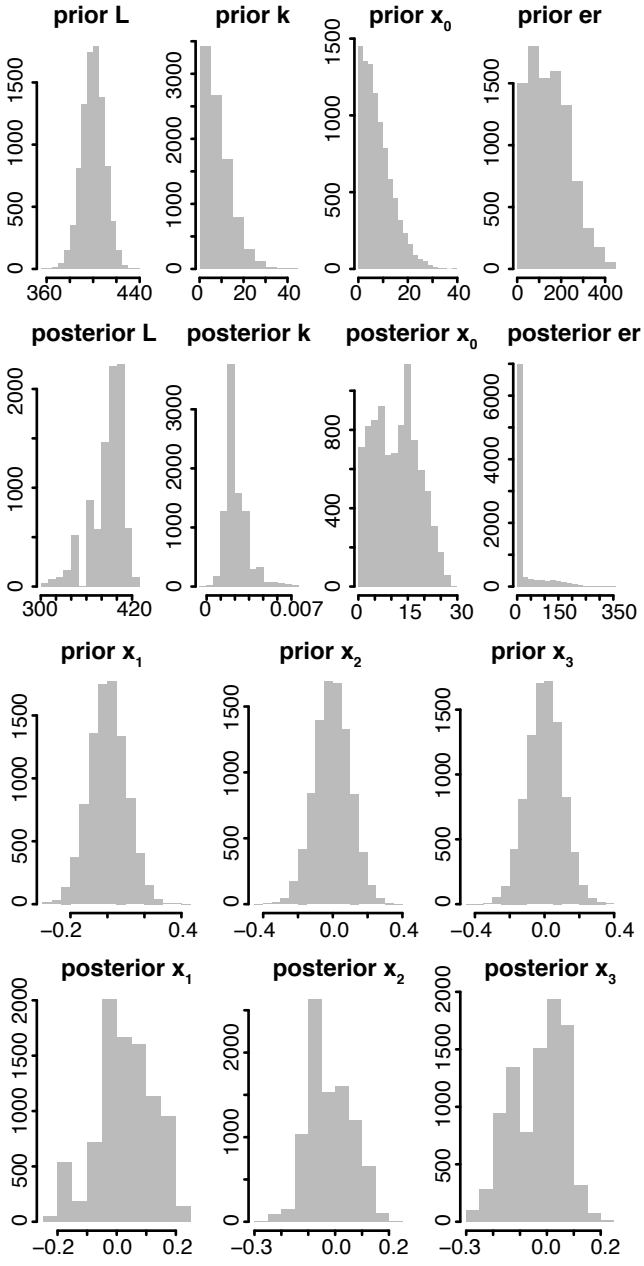

Figure 4.8: Distributions of prior and posterior values for each parameter in the model. Top row – prior distributions for L, k, x0 and er. Second row – posterior distributions for L, k, x0 and er. Third row – prior distributions for x_1, x_2 and x_3. Final row – posterior distributions for x_1, x_2 and x_3.

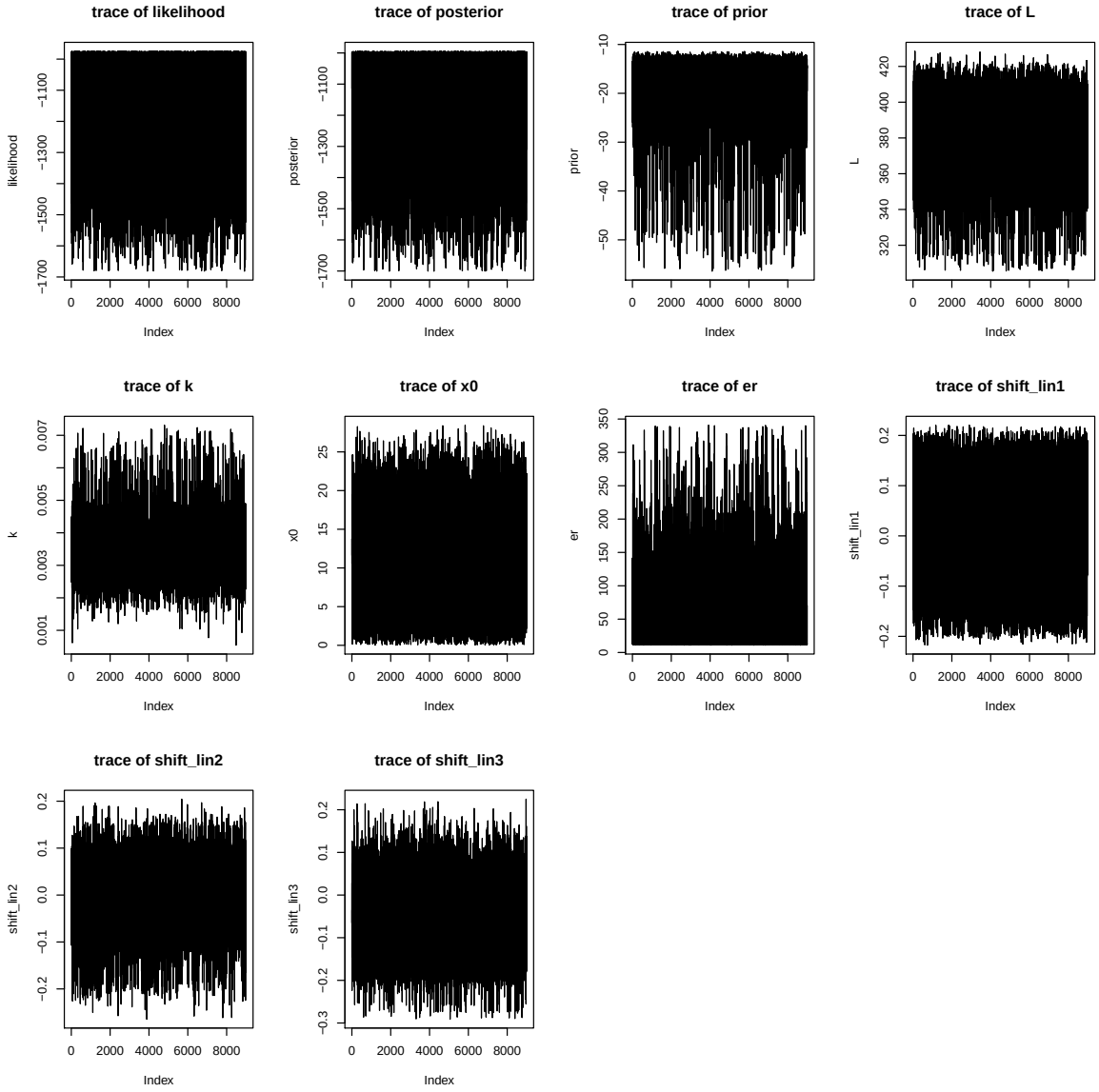

Figure 4.9: Trace plots for all parameters in the model.

4.3.3 Distribution of IS in the *S. sonnei* genome

This study aimed to asses the contribution of IS to functional inactivation within *S. sonnei*. RAST was used for functional annotation of the IS-free *S. sonnei* reference genome, which was only able to assign 57% of genes to biological systems and subsystems (see Methods section 4.2.1). Spatial distribution of IS insertion sites within the *S. sonnei* genome was investigated, which showed a non-random distribution, with various hotspots of insertion (grey shaded regions, Figure 4.10). Previous studies have noted that transposases and genes transferred via HGT can cluster around the origin of replication[349]. However, the density of IS within 10 kbp of the origin or terminus of replication was not significantly higher than any other 10 kbp region in the genome ($p > 0.8$, parametric test using normal distribution, Figure 4.10).

Next, the IS insertion data was examined to determine the contribution of IS to gene inactivation within *S. sonnei*. To explore this, the distribution of IS in coding and non-coding regions was evaluated. IS insertions were much rarer in protein-coding sequences than in intergenic regions (0.016% of genic bases vs 0.083% intergenic bases; ratio=0.19; pairwise proportion test, $p<2x10^{-16}$), consistent with the purging of genic insertions by purifying selection.

Despite this selection against genic insertions, 452 genes (10% of all protein coding sequences) were identified as interrupted by one or more IS insertions in one or more genomes. There were 59 genes that were inactivated by IS in greater than 90% of *S. sonnei* genomes (Appendix A, Supplementary Table 1). Eighteen of these 59 genes (30%) had no functional annotation. Amongst the remaining 41 genes, common functions included membrane proteins, transporters, phage, LysR and LuxR transcriptional regulators and fimbriae/flagella genes. The majority of these regions were found to be hotspots of IS insertion (grey shaded regions, Figure 4.10).

To determine if genes inactivated by IS were enriched for particular functional categories, odds ratios were calculated for each RAST system (Methods section 4.2.8). IS interruption was more frequent amongst genes not assigned to a system, and thus had unknown function (15% vs 6.7%; OR 2, 95% CI [1.7-2.5], FDR adjusted $p<10^{-11}$), and also with genes involved in monosaccharide synthesis (OR 2.3, 95% CI [1.26 – 4.0], FDR adjusted p=0.004), polysaccharide synthesis (OR 13.1, 95% CI [2.0 – 68.3], FDR adjusted p=0.004), type V and type VII protein secretion systems (OR 16.3, 95% CI [6.7 – 38.7], FDR adjusted $p=5x10^{-9}$), and regulation and cell signalling (OR 5.4, 95% CI [3.0 – 9.1], FDR adjusted $p=3x10^{-8}$).

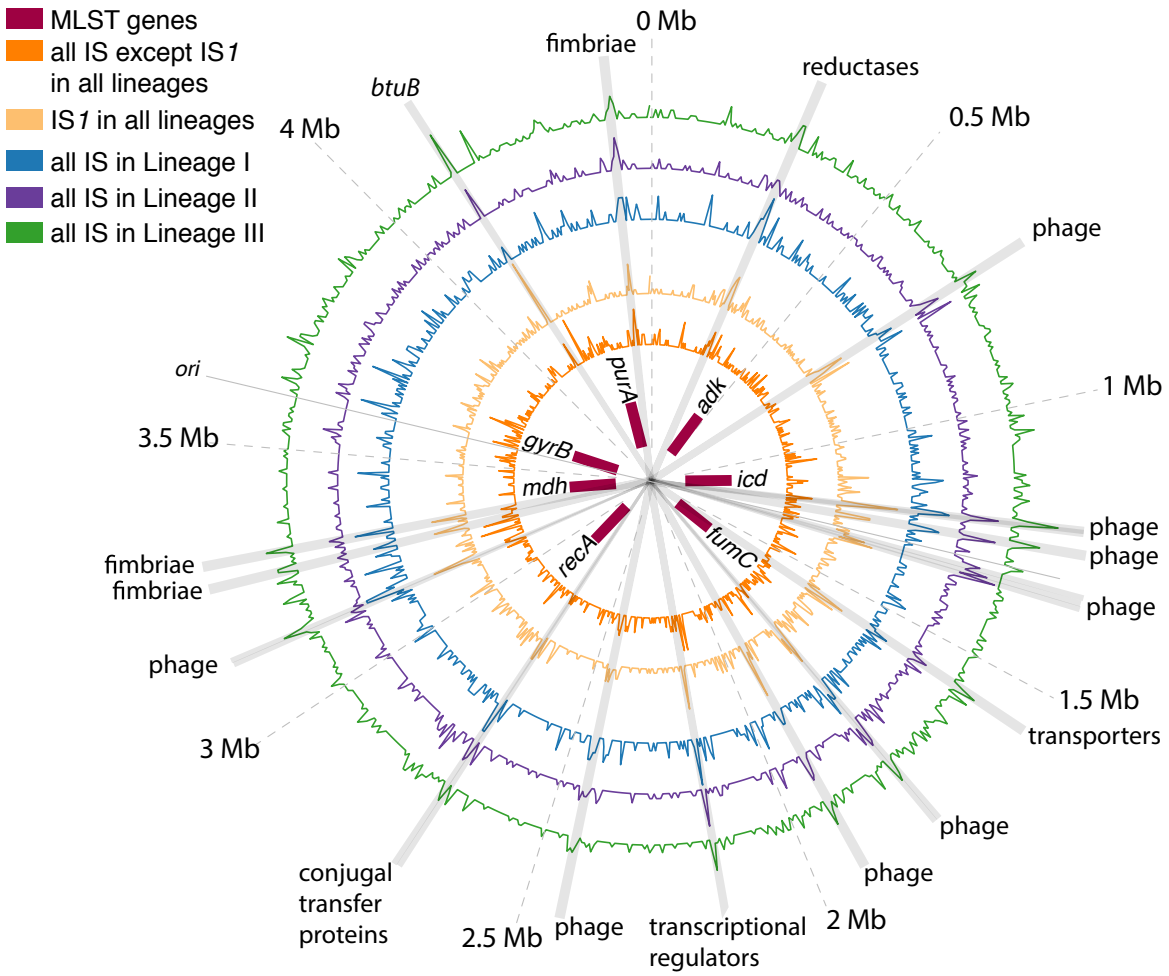

Figure 4.10: Density of IS insertion sites around the *S. sonnei* 53G chromosome. Circular map of the 53G chromosome. From outer-most to inner-most ring: IS density in lineage III; IS density in lineage II; IS density in lineage I; density of IS*1* in all lineages; density of all IS except IS*1* in all lineages; MLST genes. Grey shading indicate IS hotspot regions, annotated with the coding sequences within these regions.

4.3.4 The role of IS in negative and balancing selection amongst *S. sonnei* genomes

The enrichment of IS interruption in certain functional groups could be explained by negative selection (i.e. selection for maintenance of inactivating mutations when they arise, due to some fitness benefit associated with loss of these functions) or simply a lack of purifying selection (i.e. neutral evolution in dispensable biological systems, as opposed to purifying selection which protects against loss of essential functions). Additionally, genes enriched for IS interruption may already be pseudogenes that are undergoing further degradation by IS activity. This section describes two different approaches used to identify genes inactivated by IS that were under selection.

To detect genes that carry by high numbers of IS resulting in convergent functional gene loss, I screened for *S. sonnei* genes that were inactivated by independent IS insertions in different genomes (excluding those that were already inactivated by IS insertion or nonsense mutations, in which IS insertion represents degradation of pseudogenes rather than selective inactivation of gene function). Thirty-one genes were identified that were likely under negative or balancing selection (Appendix A, Supplementary Table 1), one of which was a transcriptional regulator, *yjjQ* (SSON53G_RS26020), which had a single, conserved interruption in all lineage III genomes, and was interrupted in a single lineage I genome. Interruption of *yjjQ* has been shown to attenuate virulence in avian pathogenic *E. coli*[350], and this regulator is part of an operon of regulators that includes *bglJ*, which regulates the expression of aryl-Beta, D-glucoside[351]. The interruption of either of these genes prevents the expression of this operon[351]. Recently, *yjjQ* has been shown to act as a transcriptional repressor of the flagellar operon *flhDC* in *E. coli*, in addition to several other virulence-associated genes[352]. As *Shigella* is non-motile, due to inactive flagella genes, it is possible that *yjjQ* is no longer required. Another of the 31 genes, *ydiM* (SSON53G_RS08475), is known to be involved in metabolism and is a transporter that helps protect cells against isoprenol toxicity[353]. The interruption of *ydiM* was mostly conserved within the global III clade of lineage III (60/77 genomes), with interruptions in two lineage I genomes, and so is likely under negative selection.

To distinguish between the possibilities of negative selection or lack of purifying selection, genic insertion rates were compared to intergenic insertions, which were assumed to be approximately neutral (Figure 4.11a). The median genic insertion rate is similar to the intergenic insertion rate

(Figure 4.11a). Interruption of genes by IS did not appear to be biased by the gene location - inactivated genes were distributed across the whole genome (Figure 4.11b). Genes interrupted by 2-fold more than the intergenic rate fall outside the majority of the distribution, suggesting that these genes may be enriched for IS interruption. However, some genes that were highly enriched for interruption (insertions per base >0.01) simply had a high rate of insertion due to their short gene length (Figure 4.11b).

To identify genes under negative selection based on their IS insertion rate, while controlling for gene length, a Poisson test was used to identify genes with a significantly higher rate of insertion than the intergenic rate, which I assumed to be roughly neutral (see Methods 4.2.7). Three genes were identified in this analysis as under negative selection - SSON53G_RS04090, a phage tail protein (7 insertions), SSON53G_RS18825, a fimbrial usher protein (11 insertions) and *btuB*, a vitamin B12 receptor (16 insertions).

Across both approaches, *btuB* was found to be a hotspot for IS insertion. There were 20 genomes containing unique IS interruptions either within *btuB* or its promoter (Figure 4.12). An additional seven genomes contained nonsense mutations within *btuB*, giving a total of 27 genomes with inactivating mutations in *btuB* or its promoter (Figure 4.12). This gene encodes a vitamin B12 receptor, which can also function as a receptor for the BF23 bacteriophage and the bacterial toxin colicin E. Colicin E is usually carried on a small plasmid, together with an immunity gene that protects the host cell against the toxin. There are nine subtypes of colicin E, labelled E1 through E9 - all bind to the same receptor, however, they differ in the way they disrupt cell function[354]. Colicin E is common amongst *E. coli* / *Shigella*, and the inactivation of *btuB* has been shown to protect *S. sonnei* cells lacking the immunity gene against the bactericidal effects of colicin E[355]. Colicin E1 and its immunity gene were identified in 46% of genomes, whilst colicin E3 and its immunity gene were identified in 12% of genomes (see Methods section 4.2.9). No other colicin E types were found.

As none of the *btuB* inactivations were conserved, I hypothesised that *btuB* was under balancing selection rather than negative selection, losing its function when exposed to colicin E but reverting to wildtype when no longer exposed in order to retain vitamin B12 receptor activity which is important for growth. Of the 27 genomes with a *btuB* inactivation, the majority (n=20, 74%) carried a colicin E and immunity gene, compared to 78% of genomes with an intact *btuB* gene. I then tested for associations between genomes carrying a specific colicin E subtype and *btuB* interruptions. Contrary to expectation, *btuB* inactivation was

more common amongst genomes that carried the colicin E1 immunity gene (33%) than those that lacked these genes (13%, OR=3.32, 95% CI [1.26 – 9.31], p=0.008), and none of the 12 genomes that carried a colicin E3 immunity gene had a *btuB* inactivation.

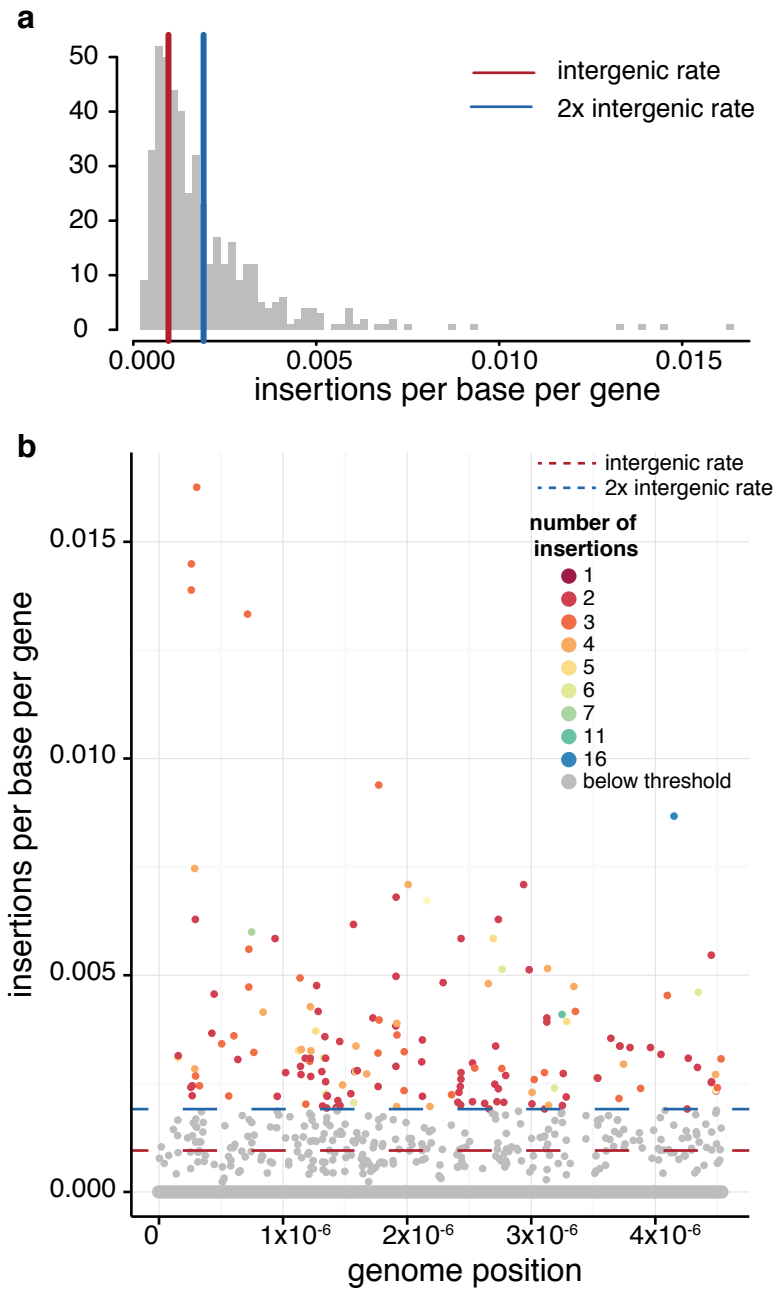

Figure 4.11: Intergenic IS insertions in *S. sonnei*. **a,** Histogram of number of insertions, per base, per gene. Intergenic rate of insertion is shown by the red line, 2x intergenic rate of insertion is shown by the blue line. **b,** Insertions per base, per gene for each gene in the *S. sonnei* genome, arranged by genome position on the x axis. Grey dots indicate genes interrupted by IS at less than the 2x intergenic rate (dashed blue line) shown in panel **a**. All other genes above the 2x intergenic rate are coloured by the number of IS insertions found within them.

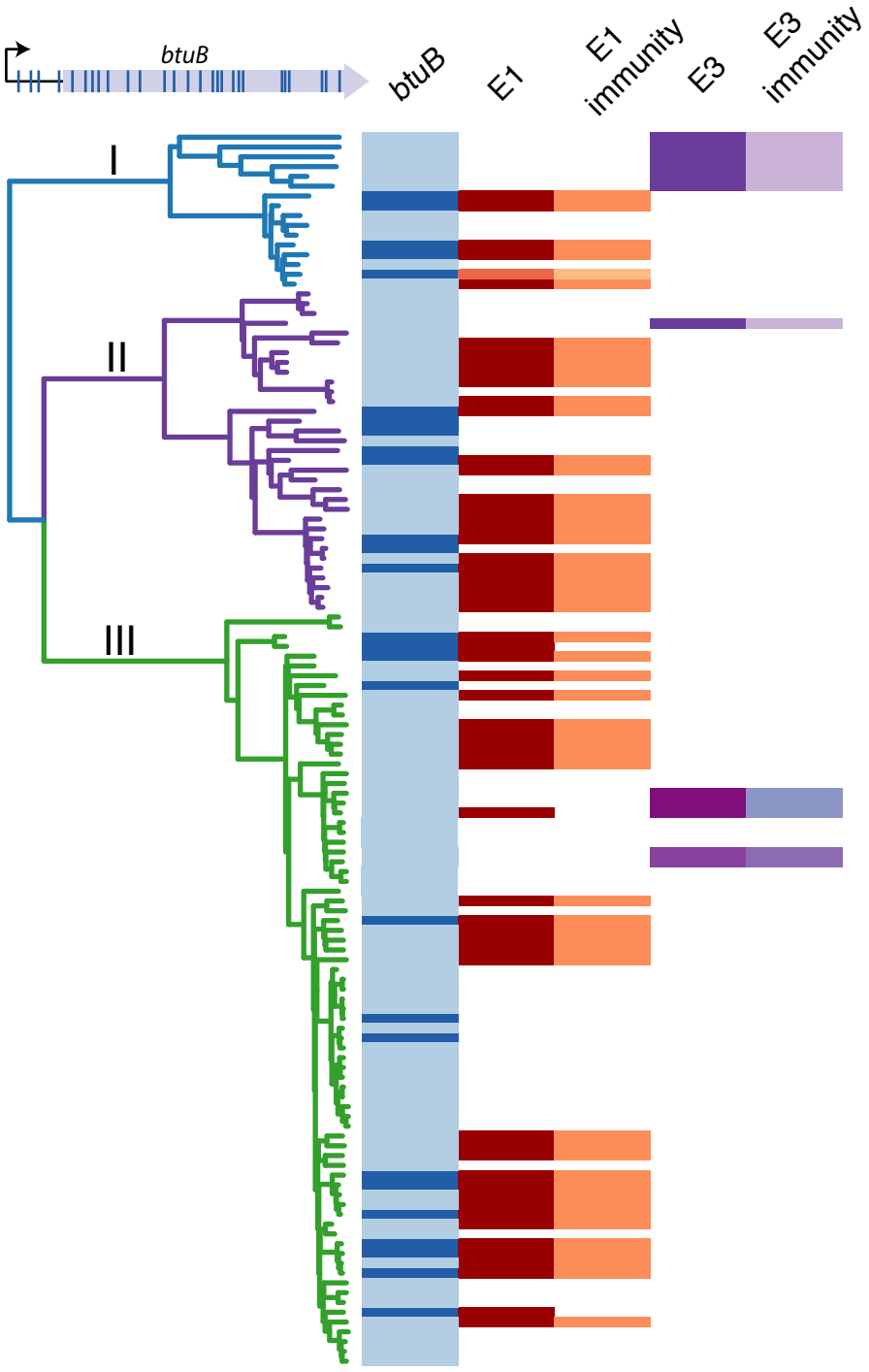

Figure 4.12: Comparison of *btuB* interruptions and presence of colicin in each *S. sonnei* genome. Light blue bars indicate an uninterrupted copy of *btuB*, dark blue bars show presence of an interruption in *btuB* or its promoter by either IS or mutation. Locations of *btuB* interruptions are shown in cartoon above phylogeny, with the promoter marked as a black arrow. Red and orange bars indicate presence of the colicin E1 toxin or immunity gene. Purple bars indicate the presence of the colicin E3 toxin or immunity gene.

4.3.5 Diversification of gene inactivation amongst *S. sonnei* lineages

I next explored the impact of IS dynamics in diversification of the three *S. sonnei* lineages. I examined the rates of genic and intergenic IS insertion across lineages. The rates of genic and intergenic IS insertion differed between lineages. Lineage III a significantly higher rate of insertion in genic regions than the other lineages (0.0017% genic bases, compared to 0.0015% in lineage II and 0.0014% in lineage I, $p=9.7x10^{-15}$, Kruskal-Wallis test), suggesting that selection against genic insertions may be weaker in lineage III than lineages I or II. Rates of intergenic insertion also significantly differed across all three lineages ($p=8x10^{-13}$, 0.026% intergenic bases in lineage I, compared to 0.034% in lineage II and 0.032% in lineage III, Kruskal-Wallis test).

Functional diversification of lineages through gene inactivation has not occurred just through IS insertion, but also through inactivation by indels or nonsense SNPs. A total of 868 genes were identified as interrupted across the set of *S. sonnei* genomes through either IS inactivation, mutational inactivation, or both. Of these 868 genes, 329 were interrupted by IS only, 416 were interrupted by mutation only, and 123 were interrupted by both. The number of conserved inactivations (inactivated in >90% of all *S. sonnei* genomes) caused by IS only or mutation only were very similar (n=16 and n=19, respectively). Twenty-four inactivations caused by either IS or mutation were conserved. These results suggest that IS inactivation and mutational inactivation perform similar roles in the generation of pseudogenes in *S. sonnei*.

Overall, there were 59 inactivated genes conserved in >90% of *S. sonnei* genomes. All lineages had a similar genome size of (median 4.6 Mbp, range 4.3 - 4.9 Mbp, see Methods 4.2.3). In addition to these 59 conserved inactivated genes, within lineage I, a further 17 genes were inactivated amongst all lineage I genomes (Appendix A, Supplementary Table 1). Lineage II contained 27 additional genes which were inactivated in all lineage II genomes, and lineage III had an additional 36 genes that were inactivated in all lineage III genomes (Appendix A, Supplementary Table 1). In all three lineages, genes with conserved inactivations were frequently also inactivated in > 80% of genomes in other lineages.

Lineage III had the greatest number of inactivated genes per genome (median 161), significantly higher than those of lineages I or II (median 133, $p=2.41x10^{-6}$; median 114, $p<2.2x10^{-16}$; respectively, using Wilcox test, Figure 4.13a). Some inactivated genes were characteristic of a particular lineage. In lineage I, six genes were inactivated in all lineage I

genomes, with few inactivations in lineages II or III (Appendix A, Supplementary Table 1). In lineage II, only three genes were inactivated in all lineage II genomes, with few inactivations in lineages I or III (Appendix A, Supplementary Table 1). One of these genes is the toxin *yhaV*, which is part of the toxin-antitoxin pair *prlF-yhaV*[356]. Expression of the *yhaV* toxin without the presence of the *prlF* antitoxin causes the bacterial colony to enter a static growth phase[356]. Lineage III contains 21 genes inactivated in all lineage III genomes with few inactivations in lineage I or II genomes (Appendix A, Supplementary Table 1). Overall, 42% of inactivated genes conserved in lineage III were not inactivated in any lineage I or II genome, suggesting that lineage III has undergone more IS-mediated functional diversification than the other lineages. This is consistent with the higher rate of IS genic insertion within lineage III compared to lineages I and II.

Next I considered the impact of IS on functional differences between strains belonging to the same and different lineages, by examining strain-specific pseudogenes. From the 868 genes interrupted in *S. sonnei*, 398 (46%) of these were strain-specific. Mutational inactivation contributed more to strain-specific inactivations than IS inactivations (233 (59%) vs 165 (42%), respectively). When comparing the number of pseudogenes that differed between pairs of strains, two strains within the same lineage differed at an average of 41 pseudogenes, compared to two strains in different lineages which differed at an average of 127 pseudogenes (Figure 4.13c). Again, mutational inactivation contributed more than IS to pairwise pseudogene differences, both within and between lineages (Figure 4.13c). This suggests that mutational inactivation contributes more than inactivation by IS in all genomes, and each is still undergoing functional diversification.

The history of pseudogene formation in the *S. sonnei* population was reconstructed and compared to the contribution of IS with other mutations. During the evolution of *S. sonnei*, it was inferred that the number of pseudogenes has increased dramatically, with all lineages seeing an increase in the number of pseudogenes compared to their mrca (Figure 4.13b, Figure 4.13d). Within lineages I and II, IS inactivations, mutational inactivations, and genes carrying both types of mutations were roughly equal amongst all extant genomes (median 43, 49 and 41 respectively for lineage I; median 38, 27 and 39 respectively for lineage II) (Figure 4.13a, Figure 4.13d). Amongst lineage III genomes, mutation has contributed significantly more to inactivation than IS (median 66 vs 44 respectively, using Wilcox test, $p<2x10^{-16}$) (Figure 4.13d). Overall, gene inactivation within *S. sonnei* has increased over time, especially

within lineage III genomes.

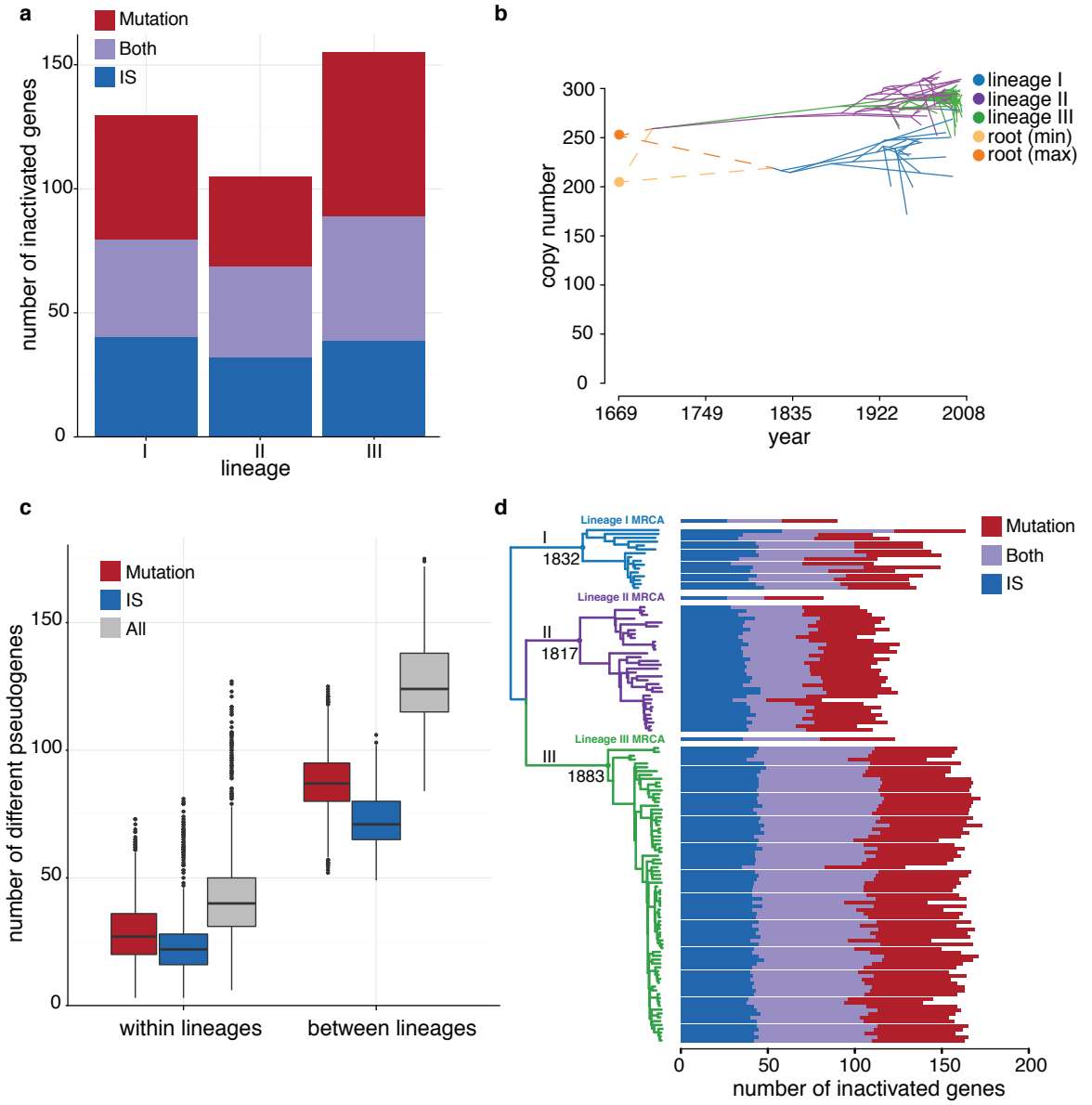

Figure 4.13: Accumulation of inactivated genes in each *S. sonnei* lineage. **a,** Bar plot with total height of bar illustrating the median number of inactivated genes in each lineage. Blue, number of genes inactivated by IS; purple, number of genes inactivated by both IS and mutation; red, number of genes inactivated by mutation. **b,** Phenogram showing total number of genes inactivated (by either IS or mutation) at each node, with branches coloured by lineage. Dark orange circle and dashed lines shows the maximum number of genes that could have been interrupted in the mrca. Light orange circle and dashed lines shows the minimum number of genes that could have been interrupted in the mrca. **c,** Box plots showing differences in pseudogenes between pairs of strains from within the same lineage and between lineages, broken down by mutation type. **d,** Number of inactivated genes in each genome. Left, phylogenetic tree of *S. sonnei*, with branches coloured by lineage. Right, bar plot showing total number of inactivated genes in each genome, coloured by inactivation type as per legend.

4.4 Discussion

4.4.1 Strengths and limitations

This study showcases how ISMapper can be utilised to examine the impact of IS on the evolution of bacterial genomes. By merging the population structure of *S. sonnei* with the presence or absence of IS insertion sites around the genome, this chapter was able to:

i) examine the differences in IS burden between the lineages of *S. sonnei*;

ii) reconstruct the history of IS acquisition and loss in *S. sonnei*;

iii) model the accumulation of IS burden across *S. sonnei*; and

iv) investigate the impact of IS on gene inactivation and pseudogene formation in *S. sonnei*, and their contribution to lineage diversification.

However, there are some important limitations with this approach. ISMapper uses a reference based approach to identify IS insertion sites, and so can underestimate the total number of IS in a query genome. IS that are in a region of the genome that is not present in the reference genome will not be detected. Additionally, the insertion of IS within other IS will also not be detected. Therefore, the analysis presented is likely to systematically underestimate in the overall IS burden found within *S. soneni* genomes, although not in ways that impact on the interpretation of genome function and selection. The majority of tools presented in Chapter 2 use a mapping approach, similar to ISMapper, and so would also underestimate overall IS burden. An assembly based approach may be able to overcome the limitation of detecting IS in regions that are not present in the reference genome, however will still struggle to detect IS that have inserted within other IS, as these regions are inherently difficult to assemble. Long read sequencing is likely the only method able to overcome both limitations.

The *S. sonnei* dataset used in this study contained representatives from the three major lineages discovered by Holt *et al.*[250], with a diverse geographical and temporal distribution. However, there are fewer lineage I genomes (n=16) compared to lineage II and III (n=33 and 77, respectively), and the lineage I genomes were more distantly related to each other than lineage II or III genomes. Greater sampling density of lineage I genomes would aid comparisons between this lineage and the others.

Within this dataset, there were few *S. sonnei* genomes that included sequence from the virulence plasmid, pINV, in their genomes, at an appropriate level of coverage and depth for study using ISMapper. The virulence plasmid is often lost during growth on solid media, making it impossible to sequence[357]. As the virulence plasmid was missing from 70% of strains in this dataset, investigation of IS insertion sites on the plasmid was not possible. Given the differences in chromosomal IS content between lineages in *S. sonnei*, if full plasmid sequence was available for all genomes in this dataset, I would expect to see similar differences in IS content amongst the plasmids within each lineage. Future investigation could also examine the contribution of IS and mutation to the formation of pseudogenes within the virulence plasmid.

In this chapter, I attempted to determine the theoretical maximum IS load for *S. sonnei*, based on the inferred past trajectory of IS activity and assuming a logistic growth model could explain IS expansion in the *S. sonnei* genomes. This model predicted an IS saturation point of 398 IS copies (section 4.3.2). In this model, the three lineages were assumed to be at different points on the logistic curve, based on the results of the linear regression and differences in gain rates (section 4.3.1). Lineage I was assumed to be on the steeper section of the curve, as the extant lineage I genomes had a steeper regression curve and similar copy number to the mrca of lineages II and III. The placement of the points on the x-axis was allowed to vary under the model, however tight priors were placed on these values to prevent non-identifiability: in this Bayesian framework, there were two possible ways to improve the fit of the model - either by shifting the points on the x-axis, or adjusting the parameters of the curve. If there were not strong priors on the parameters controlling the location of the points on the x-axis, then the model would likely run into a plateau of high likelihoods, where there would be infinite combinations of point locations and line parameters, all with an equally good fit. The posterior distribution of the parameters for the location of the points on the x-axis was different to their prior distributions, indicating that the tight priors on these parameters were not overly informative or driving the results (Figure 4.8).

4.4.2 IS in *S. sonnei* behave differently depending on lineage

The data present here indicate that all three lineages of *S. sonnei* are still undergoing IS expansion, albeit at different rates. Each lineage is accumulating IS over time, as evidenced by the linear regression analysis on extant copy number (section 4.3.1) and the modelling of inferred copy

number at the internal nodes of the tree (section 4.3.2). Additionally, the presence of strain-specific insertions indicates that transposition activity is still occurring in all genomes (with the exepction of IS*609*, section 4.3.1). IS burden in contemporary genomes suggests that lineage I may be currently accumulating IS faster than lineages II or III. However, combining the inferred ancestral IS copy numbers with those numbers observed in contemporary isolates, indicates that either:

i) the mrca of lineages II and III underwent rapid accumulation of IS during the 18th century (acquiring 55 insertions in 35 years), followed by slower increases in each lineage in recent years, resulting in a current IS load of 300 IS per genome, while lineage I has undergone a much more steady increase in IS copy number since its mrca (Figure 4.14a); or

ii) lineages II and III have undergone a fairly steady increase in IS burden since the *S. sonnei* mrca (8 IS in 35 years), while the mrca of lineage I lost significant numbers of IS since the *S. sonnei* ancestor, before returning to a steady increase in copy number (Figure 4.14b).

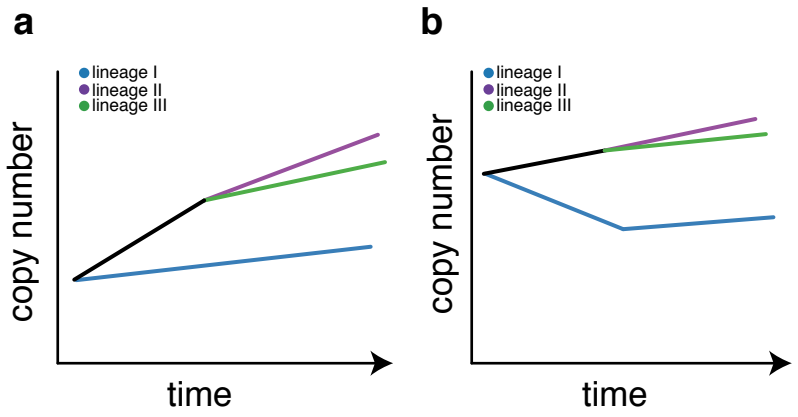

Figure 4.14: Two different hypotheses for the burden of IS in the mrca of *S. sonnei*. a, Lineages II and III rapidly accumulate IS, while lineage I accumulates IS at a steadier rate. **b,** Lineages II and III accumulate IS at a steady rate, while lineage I rapidly loses IS, then gains additional IS copies later at a more steady rate.

It is difficult to tease apart these two hypotheses as the inference of IS burden at the mrca of all *S. sonnei* is not possible with the current data, due to the inability to perform ancestral reconstruction of IS insertion sites that are found only in the lineage I mrca, but not the mrca of lineages II and III, and vice versa. Either these sites were present at the root and lost in the lineage I mrca, or 55 of these sites were absent at the root and gained in the mrca of lineages II and III, or more likely, there was some combination of these two scenarios.

Supporting hypothesis i) were the high proportion of IS*1* sites in the mrca of lineages II and III. The proportion of IS*1* sites present in the mrca of lineages II and III, but absent in the lineage I mrca, was higher than the proportion of IS*1* sites present in the lineage I mrca, but absent from the mrca of lineages II and III. IS*1* was consistently the top IS on all metrics used to measure differences in IS behaviour amongst the twelve IS investigated in this dataset (section 4.3.1). IS*1* had the highest number of strain-specific insertions (section 4.3.1), and also the highest number of independent gains across the tree (section 4.3.2), suggesting that IS*1* has a high transposition activity in *S. sonnei*.

There are two mechanisms that could be responsible for a rapid loss of IS in the lineage I mrca (hypothesis ii)). Firstly, the lineage I mrca may have undergone large genomic deletions, resulting in the loss of several IS. If this were the case, lineage I genomes would be smaller than lineage II or III genomes. However, all lineages had roughly the same size genome (section 4.3.5). Additionally, the majority of inactivated genes conserved in lineage I were also inactivated in >80% of lineage II and III genomes (section 4.3.5), indicating that lineage I is not missing large numbers of genes present in lineages II or III. Secondly, IS that transpose via a cut and paste mechanism, where the IS is excised and moved, rather than copied, may have resulted in fewer IS in the lineage I mrca if these IS migrated to a different replicon (for example, the invasion plasmid). Within *S. sonnei*, only IS*4* is known to transpose via a cut and paste mechanism. However, the proportion of IS*4* in the lineage I mrca is similar to the proportion found in the lineage II/III mrca (section 4.3.2), making this hypothesis unlikely.

Given all these scenarios for both hypotheses, biologically, it seems more likely that the mrca of lineages II and III rapidly gained IS, rather than lineage I rapidly losing IS. This may have been caused by some relaxation of control over transposition for IS*1* within the lineage II/III mrca that was not present in the mrca of lineage I.

4.4.3 IS-mediated genome decay is ongoing in *S. sonnei*

It has previously been suggested that functional gene loss was an important step in the adaptation of *Shigella* to the human host, with metabolic genes, motility genes, membrane genes and transporters often decayed or interrupted in all four species[358]. IS expansion played a significant role in the evolution of other pathogens. In *Y. pestis*, 34% of pseudogenes were caused by IS-mediated gene interruptions[60]. IS-mediated gene loss was a major evolutionary

process that contributed to the evolution of *B. mallei* from *B. pseudomallei*[53]. In all of these examples, IS have contributed to genome decay, where gene inactivation by IS has created large numbers of pseudogenes. These data indicate that IS have played a substantial role in functional gene loss in *S. sonnei*. Of the 868 genes inactivated in at least one *S. sonnei* genome, 329 were IS-mediated inactivations, with a further 123 inactivations mediated by both IS or mutation (section 4.3.3).

There was likely a recent evolutionary bottleneck in the mrca of *S. sonnei* associated with becoming restricted to the human host. These data suggest that all 12 IS were already present at the time of this bottleneck and have become fixed in the *S. sonnei* population. All lineages appear to be continuing to undergo genome decay, with an increase in the number of pseudogenes since the *S. sonnei* mrca. Most notably, lineage III appears to be undergoing the most genome decay, with significantly more inactivated genes than either lineages I or II.

Several categories associated with sugar synthesis were found to be inactivated in *S. sonnei* (section 4.3.4), which may contribute to structure of the outer membrane presented to the environment. Enrichment for inactivation of these genes suggests that there may be antigenic variation occurring, driven by selection from the host immune system. This phenomenon has been observed in several bacteria, where changes to the sugar molecules expressed on the surface of their cells leads to evasion of the human immune system[68,70,273].

Loss of function may confer a selective advantage, since loss of fimbriae is known to be an important step in the evolution of *Shigella* species from other *E. coli*[341,359], and prophage suppression can be beneficial to bacterial hosts[94]. Functional gene loss does not always need to be permanent, as was shown with the frequent but transient convergent inactivation of *btuB* (section 4.3.4). My results indicate that the lack of a colicin immunity gene may not be the selective pressure behind *btuB* inactivation. However, there was some positive association between the presence of colicin E1 and a colicin E1 immunity gene and an inactive *btuB* gene. This may be due to an ineffective colicin E1 immunity gene in some strains, and so these strains still require an inactive *btuB* receptor. However, *btuB* is also a receptor for the phage BF23[360], and this may explain the high number of genomes within this data set containing inactivated *btuB* genes. Lack of conservation of *btuB*-inactivating IS insertions or mutations in *S. sonnei* (each individual interruption affecting *btuB* was observed in just one isolate and not passed on, Figure 4.12) suggests that whatever selective pressure is present is temporary, and is likely balanced by positive selection for *btuB* activity in the absence of exposure to colicins or

phage, as vitamin B12 is ultimately essential for long-term growth.

4.5 Summary

Overall, this chapter provides a novel framework for the investigation of the influence of IS in the evolution of bacterial pathogens, and reveals the important role that IS have played in the evolution of *S. sonnei*. Examination of IS in *S. sonnei* highlighted the large IS burden in all *S. sonnei* genomes, which modelling suggested is still increasing towards an overall load of 398 IS per genome. IS burden differed between lineages, with lineages II and III carrying higher IS loads compared to lineage I. The role of IS in pseudogene formation and genome decay was investigated, and demonstrated that IS were still playing a role in pseudogene formation. However, mutational processes are still responsible for a high proportion of gene inactivation. Each lineage in *S. sonnei* were found to be still undergoing genome decay, through IS-mediated and mutation gene inactivation.

Chapter 5

Insertion sequences in *Shigella dysenteriae* and *Shigella flexneri*

5.1 Introduction

This chapter aims to investigate the dynamics of IS in *S. dysenteriae* and *S. flexneri* genomes, and compare these dynamics to *S. sonnei*. As discussed in Chapter 1, *S. dysenteriae* and *S. flexneri* are two patho-adapted lineages of *E. coli* that cause dysentry or shigellosis[202]. A fourth *Shigella* species, *S. boydii*, is found predominantly on the Indian sub-continent[202]. *S. boydii* is rarely found outside Bangladesh, and with only 28 genomes available from a study in 2016[361], this species is not analysed in depth in this thesis.

Each *Shigella* species has undergone genome reduction compared to *E. coli*, but each species is at different points in their evolutionary path. *S. dysenteriae* has the most reduced genome (3380 genes in a 3.84 Mbp genome, excluding IS), and is mostly associated with epidemics of dysentery. *S. flexneri* genomes are much larger than *S. dysenteriae*. *S. flexneri* 2a strain 301 has 3905 genes and a 4.28 Mbp genome, excluding IS. *S. flexneri* usually causes more endemic disease in developing nations. In some regions of the world, *S. flexneri* is gradually being replaced by *S. sonnei*[251], which has the largest genome (4445 genes, in a 4.56 Mbp genome, excluding IS). *S. boydii* has a similar genome size to *S. flexneri* (4.52 - 4.62 Mbp, including IS).

Previous studies have already described the population structure of *S. dysenteriae* and *S. flexneri*, which was presented in Chapter 1. The genomes from these studies of *S. dysenteriae* and *S. flexneri* are used in this chapter to investigate IS dynamics within these two species. The population study of *S. dysenteriae* investigated 325 genomes of *S. dysenteriae* Sd1, as serotype Sd1 is the causative agent of all the major dysentery outbreaks since the late 19th century. The 325 *S. dysenteriae* Sd1 genomes were isolated from 66 countries, spanning the years 1915 - 2011, and included 14 isolates obtained during the First World War[260]. The phylogeny of these genomes revealed that *S. dysenteriae* consisted of four distinct lineages[260]. I was responsible for performing BEAST analysis to date the emergence of this species (published in *Nature Microbiology*, 2016, Njamkepo *et. al.*[260]). I used a smaller subset of 125 genomes, as the full dataset was too large to reach convergence in BEAST. Based on the analysis of the smaller Sd1 subset, I estimated that the mrca of *S. dysenteriae* Sd1 existed circa 1747 (95% HPD 1645 - 1822), and has since diversified into four lineages, three of which are globally disseminated (Lineage I was represented by just one isolate from England, and was separated by 1,200 SNPs from the rest of the tree)[260]. Across the subset of 125 genomes the mean nucleotide divergence between pairs of strains was 0.013%. The mrca of lineages II and III both existed in the late

19th century - lineage II in 1877 (95% HPD 1865 - 1888) and lineage III in 1889 (95% HPD 1881 - 1897)[260]. Lineage IV is actually a subclade of lineage III that arose in 1929 (95% HPD 1918 - 1939)[260]. The same subset of 125 *S. dysenteriae* genomes are used in this chapter to investigate IS dynamics in *S. dysenteriae*.

The largest comparative genomics study to date of *S. flexneri* is a study of 351 *S. flexneri* genomes from Africa, Asia, Central and South America, and historical isolates from North America and Europe, spanning the years 1914 - 2011 (Connor *et. al.*[264]). This study showed that *S. flexneri* is the most diverse *Shigella* species, consisting of seven distinct lineages. Across all genomes, mean pairwise nucleotide divergence was 0.07%[264]. Each of the seven lineages were found to have a similar level of diversity to the entire *S. sonnei* species, with a mean pairwise nucleotide divergence of 0.012%[264]. Due to the diversity found in *S. flexneri*, each lineage was dated separately, and no estimate was obtained for the mrca of all *S. flexneri*. Dates for the mrca of each lineage varied. Lineages 1, 2, 4 and 6 were the oldest, with their mrca existing between 1341 and 1649[264]. Lineages 3 and 5 were more recent, with mrca's in the early 19th century[264]. Each lineage contains at least one genome collected since 2008[264]. All seven lineages were present on multiple continents, and contained representatives from multiple serotypes and multiple geographic locations, suggesting that there has been long-term colonisation of lineages that co-exist with other lineages in the same geographic contexts[264]. All 351 genomes in the Connor *et. al.*[264] study are used in this chapter to investigate IS dynamics in *S. flexneri*.

5.1.1 Aims

This chapter explores the IS dynamics within *S. dysenteriae* and *S. flexneri* using a subset of 125 *S. dysenteriae* genomes and the phylogenetic structure from Njamkepo *et. al.*[260], and all 351 *S. flexneri* genomes and phylogenetic structure from Connor *et. al.*[264]. Similar methods used for studying IS dynamics in *S. sonnei* (Chapter 4) have been applied here to both data sets.

The specific aims of this study were:

i) To investigate the burden of IS within *S. dysenteriae* Sd1, and differences in IS content and behaviour between lineages;

ii) To investigate the burden of IS within *S. flexneri*, and differences in IS content and

behaviour between lineages;

iii) To compare the behaviour of IS between three *Shigella* species;

iv) To understand the differences in IS dynamics between each *Shigella* species and the expansion of IS within *Shigella* and the rest of *E. coli*; and

v) To examine the impact of IS on genome decay and functional diversification within each *Shigella* species.

IS insertion sites were detected with ISMapper in all 125 *S. dysenteriae* genomes and all 351 *S. flexneri* genomes to examine burden within each species and each lineage within species (sections 5.3.1 and 5.3.2). The burden and behaviour of IS in *S. dysenteriae* and *S. flexneri* were compared with *S. sonnei*, revealing five IS common to all three *Shigella* species that had undergone signficant expansion (section 5.3.3). The distribution of these five common IS was investigated in a sample of 1000 *E. coli* genomes, representative of the wider *E. coli* population (section 5.3.4). To explore the dynamics of IS in pathogenic *E. coli* lineages, and compare this to the three patho-adpated *Shigella* species, IS expansion in three well known human pathogenic *E. coli* lineages was examined (section 5.3.4). To examine the role of IS in gene in activation across all three *Shigella* species, the contribution of IS and mutational inactivation was compared in *S. dysenteriae* and *S. flexneri* (sections 5.3.5.1 and 5.3.5.2). Gene inactivations in the three *Shigella* species were compared to investigate and compare the amount of genome decay ongoing in each species (section 5.3.5.3).

5.2 Methods

5.2.1 *S. dysenteriae* data and analysis

The *S. dysenteriae* data contained of 125 genomes from Njamkepo *et. al.*.[260], sequenced on the Illumina platform, generating 100 - 146 bp paired end reads, with an average read length of 115 bp. Average read depth was 193x (range = 31.8x - 2889x). IS were identified in the complete reference genome *S. dysenteriae* Sd197 (accession NC_007606) in the same manner as *S. sonnei*, to created an IS-free *S. dysenteriae* Sd197 reference genome (doi: 10.4225/49/589c305ea8b91, see section 4.2.1). Six IS were detected in the complete *S. dysenteriae* Sd197 reference genome for downstream analysis - IS*1*, IS*2*, IS*4*, IS*600*, IS*911* and ISEc8. IS insertion sites for each IS

were detected in all 125 genomes using ISMapper with default settings, against the IS-free Sd197 reference. The presence or absence of each IS site on each internal node of the phylogeny was determined using maximum parsimony ancestral state reconstruction, as described in section 4.2.5. Identification of non-synonymous SNPs and intergenic indels was conducted using the same method as section 4.2.2, using the IS-free Sd197 reference genome.

All genes in the IS-free Sd197 reference genome were assigned a functional category with RAST, using the same methods as section 4.2.1.

5.2.2 *S. flexneri* data and analysis

The *S. flexneri* dataset consisted of 351 genomes from Connor *et. al.*[264], sequenced on the Illumina HiSeq, generating 100 bp paired end reads, with an average read length of 102 bp. Average read depth was 102x (range 18.9x - 419x). IS were identified in the complete reference genome *S. flexneri* 2a strain 301 (accession AE005674), creating an IS-free *S. flexneri* 2a strain 301 reference genome (doi: 10.4225/49/589c30d1e976b, see section 4.2.1). Twelve IS were detected for downstream analysis - IS*1*, IS*2*, IS*4*, IS*600*, IS*609*, IS*911*, IS*1203*, IS*150*, ISEc17, ISEhe3, ISSlf3 and ISSfl4. Locations for each IS were detected in each of the 351 genomes using ISMapper with default settings, against the IS-free reference strain 301. The presence or absence of each IS site on each internal node of the phylogeny was determined using maximum parsimony ancestral state reconstruction, as described in section 4.2.5. Identification of non-synonymous SNPs and intergenic indels was conducted using the same method as section 4.2.2, using the IS-free reference strain 301.

All genes in the IS-free strain 301 reference genome were assigned to a functional category with RAST, using the same methods as section 4.2.1.

5.2.2.1 *S. flexneri* phylogenies

All 351 genomes were mapped to reference genome *S. flexneri* 2a strain 301 using RedDog v1b.9 to call SNPs for phylogenetic analysis. SNPs in repeat regions (defined as IS, detected in section 5.2.2, or phage, detected using PHAST[362]) were removed. The resulting alignment of 40,073 SNPs were used to construct a maximum likelihood phylogeny using RAxML v8.2.8[363],

with a GTR+Γ substitution model and ascertainment bias correction. BEAST analysis was not performed on the full dataset, due to the large number of genomes in the alignment.

5.2.3 Identification of IS in *S. boydii* genomes

Two complete *S. boydii* genomes (Sb227, accession NC_007613; CDC 3083-94, accession NC_007613) were analysed using ISSaga[342] to identify IS present in *S. boydii*. Only IS with >80% nucleotide identity that were not identified as pseudogenes or probable false positives were considered present in each genome.

5.2.4 Identifying orthologous genes in *Shigella*

Orthologous genes were determined by comparing each IS-free *Shigella* reference genome to another using the reciprocal shortest distance algorithm (https://github.com/todddeluca/reciprocal_smallest_distance)[364]. Default parameters and the BLAST v2.2.30+ were used.

5.2.5 IS*1* sequence analysis

To investigate the differences in IS*1* sequences between *Shigella* and *E. coli*, IS*1* sequences were extracted from *S. sonnei* references 53G and Sso46, *S. flexneri* 2a strain 301 and 2a 2457T, *S. dysenteriae* Sd197 and 10 *E. coli* genomes (Table 5.1). Completed genomes were used for this analysis, since the Illumina read sets are too short to resolve sequence variation between IS*1* copies. IS*1* sequences were identified in each genome by searching for the IS*1* reference (ISFinder accession M37615) using BLAST+ v2.2.30[267]. All IS*1* sequences were aligned with MUSCLE v3.8.31[365]. RAxML v8.2.8[363], with a GTR+Γ model, was used to infer a phylogenetic tree of all IS*1* sequences. Alignments of IS*1* sequences from each *Shigella* species were manually inspected to look for mutations within the *insA* and *insB* frameshift region.

Table 5.1: Genomes and accessions used for IS*1* phylogeny.

Accession	Genome	# IS1 copies	Sequencing Method	Sequencing Group
NC_016822	*S. sonnei* 53G	166	Capillary sequencing	Wellcome Trust Sanger Institute
NC_007384	*S. sonnei* Sso46	168	Capillary sequencing	Chinese Ministry of Public Health
NC_007606	*S. dysenteriae* Sd197	154	Capillary sequencing	Chinese Ministry of Public Health
AE005674	*S. flexneri* 2a str. 301	109	Capillary sequencing	Chinese Ministry of Public Health
AE014073	*S. flexneri* 2a str. 2547T	105	Capillary sequencing	University of Wisconsin-Madison
NZ_CP008957	*E. coli* O157:H7 str. EDL933	2	PacBio and Illumina hybrid assembly	University of San Diego
AP009378	*E. coli* str. SE15	3	454 and Sanger sequencing	University of Tokyo
NC_018658	*E. coli* O104:H4 str. 2011-C	10	454 and Illumina hybrid assembly	Los Alamos Public Laboratory
NC_013353	*E. coli* O103:H2 str. 12009	1	Capillary sequencing	University of Tokyo
AE014075	*E. coli* str. CTF073	1	Capillary sequencing	University of Wisconsin-Madison
CP014197	*E. coli* str. MRE600	84	PacBio with HGAP assembly	Wiell Cornell Medical College
CP016546	*E. coli* O177:H21	9	PacBio with HGAP assembly	Leiden University Medical Center
NC_008563	*E. coli* str. APEC01	2	Capillary sequencing	Iowa State University
NC_011415	*E. coli*str. SE11	1	Capillary sequencing	University of Tokyo
NC_010473	*E. coli* K12 substr. DH10B	11	Capillary sequencing	University of Wisconsin

5.2.6 *E. coli* data and analysis

1000 *E. coli* genomes from GenomeTrackr, collected by Ingle *et. al.*[366], were used as a representative sample of the *E. coli* population, as the sample of genomes contained a range of both human and animal isolates from foodborne outbreaks. Across all *E. coli*, there are ~2, 200 genes shared amongst all isolates[367]. As such, there is no close reference genome for all genomes in the population. Therefore, IS detection with a reference-based approach as implemented in ISMapper was not appropriate for this dataset. As an alternative method of estimating IS copy number, each *Shigella* or *E. coli* readset was mapped to IS*1*, IS*2*, IS*4*, IS*600* and IS*911* references and the seven loci of the Achtman MLST scheme using SRST2 v0.2.0[368]. The genomes had already been assigned to 161 eBURST groups based on MLST data by Ingle *et. al.*[366], with a single ST usually forming the majority of genomes within each eBURST group. Rare eBURST groups were defined as groups that contained less than ten genomes. These rare eBURST groups were combined together, leaving 12 major eBURST groups ($n \geq 11$ genomes) in addition to the rare groups (total of 355 genomes from 176 rare groups). To obtain an approximate copy number for each IS within each eBURST group, the mean read depth at each IS, calculated by SRST2[368] based on mapping to the reference IS sequence using bowtie2[297], was divided by the mean read depth across the MLST loci (i.e. single copy chromosomal genes), calculated in the same manner, and averaged across all genomes of the same eBURST group.

5.2.7 Detection of IS in ST131, ST11 and O104:H4 *E. coli*

To investigate whether IS expansion has occurred in other pathogenic lineages of *E. coli*, three different clonal lineages representing different pathotypes and STs of *E. coli* were obtained. ST131 uropathogenic *E. coli* (UPEC) (n=82) and reference strain SE15 (accession AP009378[369]) were collated from two separate studies[370,371]. ST11 enterohemorrhagic *E. coli* (EHEC) (n=199) were extracted from the MLST results of the 1000 GenomeTrackr genomes collected by Ingle *et. al.*[366]. The ST11 reference genome selected was *E. coli* O157:H5 strain EDL933 (accession NZ_CP008957[372]). Representatives of the German outbreak clone O104:H4 (n=36)[373] with reference *E. coli* strain 2011C-3493 (accession NC_018658) were obtained from NCBI.

ST131 reads had an average read length of 100 bp, and an average read depth of 60x (range 34x

to 202x). ST11 reads had an average read length of 170 bp, and an average read depth of 80x (range 16x to 245x). O104:H4 reads had an average read length of 98 bp, and an average read depth of 57x (range 8x to 214x).

IS were identified in each *E. coli* reference genome using the ISSaga[342], as described in section 4.2.1. This revealed five IS in ST131, nine IS in ST11 and thirteen IS in O104:H4. For each lineage, the IS detected in their reference were used as queries with ISMapper to identify the IS insertion sites in each genome of that lineage.

5.3 Results

5.3.1 IS distribution in *S. dysenteriae* Sd1

Six IS were identified in *S. dysenteriae* Sd1, belonging to four different IS families (Table 5.2). A total of 609 different IS insertion sites were detected, with a median number of 198 IS copies per genome (Figure 5.1f). All four lineages contained a similar median number of IS copies (lineage I, 188 (single genome); lineage II, 197; lineage III, 201; lineage IV, 197; Figure 5.1d), and there was little difference in the relative proportion of the various IS between lineages (Figure 5.1e-f). Overall, 161 IS insertion sites were conserved in >90% of all *S. dysenteriae* genomes. Despite similarities in IS copy number between lineages, the abundant lineages II, III and IV could be distinguished by PCA analysis of their specific IS insertion sites (PC1 - 14.7% variation, PC2 - 5.7% variation) (Figure 5.1c, Figure 5.2).

IS*1* was the most abundant IS, contributing to approximately 60% of all IS insertion sites in each genome (Figure 5.1a, Table 5.2). The next most abundant were IS*600* (20%) and IS*2* (11%), with the remaining IS, IS*4*, IS*911* and IS*Ec8* contributing to less than 5% of sites each (Figure 5.1a, Table 5.2). Only one ISEc8 insertion site was observed, which had been lost in four genomes.

Despite the collection comprising isolates spanning 100 years, there was no evidence of increased IS copy number during this time. Linear regression of isolation date and IS copy number in each genome was no evidence of a linear relationship between IS copy number and year of isolation within lineages II and III (R^2 0.7% and 3.3% respectively, $p > 0.3$, see Figure 5.3). Lineage IV showed a statistically significant relationship ($p=0.044$) between isolation date and IS copy number, with IS copy number increasing by approximately 0.13 IS insertion

sites year^{-1}, however the R^2 value indicates that this relationship explains only 6% of the variation, and this lineage has only been evolving for ~40 years (Figure 5.3). Ancestral state reconstruction inferred ~200 IS at the mrca *S. dysenteriae* and the mrca of each lineage, similar to the IS copy number found in extant genomes (Figure 5.1e).

To assess current IS activity, strain-specific IS insertions were tallied for each IS. IS*600* and IS*1* were the most active, with over 100 strain-specific positions each (Figure 5.1b, Table 5.2). IS*2* and IS*911* had similar activity levels, with 21 and 27 strain-specific sites, respectively (Table 5.2). Together with the linear regression and ancestral state reconstruction, these data indicate that IS copy number is no longer increasing in *S. dysenteriae*, however, this is not due to inactivity of the IS, but rather is balanced by both gain and loss events, consistent with having reached a maximum IS saturation (Figure 5.2, Figure 5.1e).

Table 5.2: IS detected in 125 *S. dysenteriae* genomes using ISMapper analysis.

IS	IS family	# sites across 125 genomes	mean proportion per genome	# strain-specific insertions
IS*1*	IS*1*	227	0.59	112
IS*600*	IS*3*	222	0.20	137
IS*2*	IS*3*	49	0.11	21
IS*911*	IS*3*	40	0.045	27
IS*4*	IS*4*	20	0.045	6
ISEc8	IS*66*	1	0.005	0

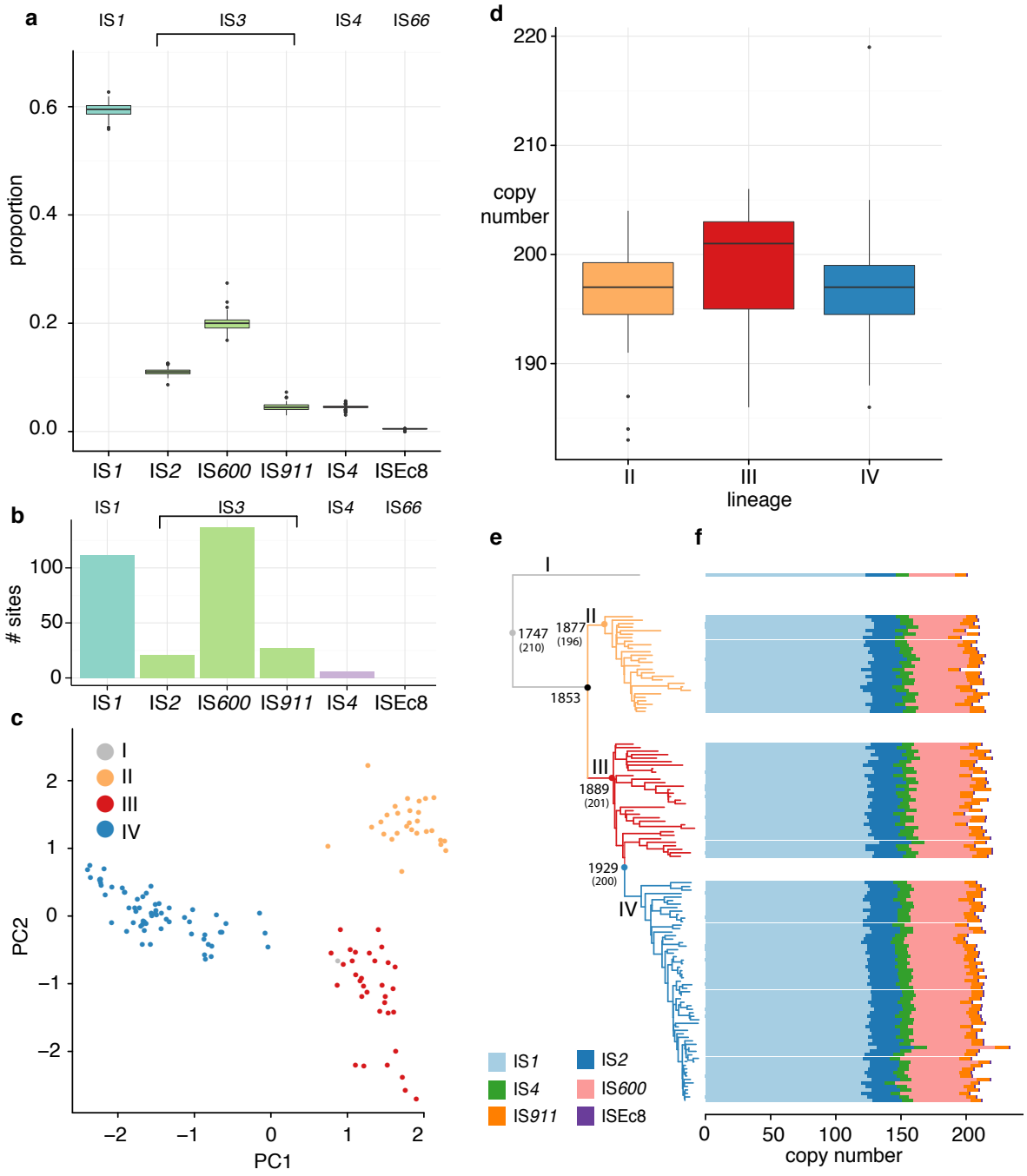

Figure 5.1: Burden of IS within *S. dysenteriae*. a, Box plots of IS proportion for each IS within the genome, with IS family indicated across top. **b,** Bar plots showing the number of strain-specific insertion sites for each IS, with IS family indicated across top. **c,** PCA of IS profiles for each genome, illustrating that lineages can be separated based on their IS profile. **d,** Box plots showing range of copy numbers within each lineage (lineage I not shown as it consists of a single isolate). **e,** Maximum clade credibility tree from Njamkepo *et. al.*[260], with branches coloured by lineage. Dates indicate time of ancestor at major nodes, numbers in brackets show inferred IS copy number at that node. **f,** Bar plots showing total copy number of each IS for each isolate in the tree, with segments coloured by IS as per legend.

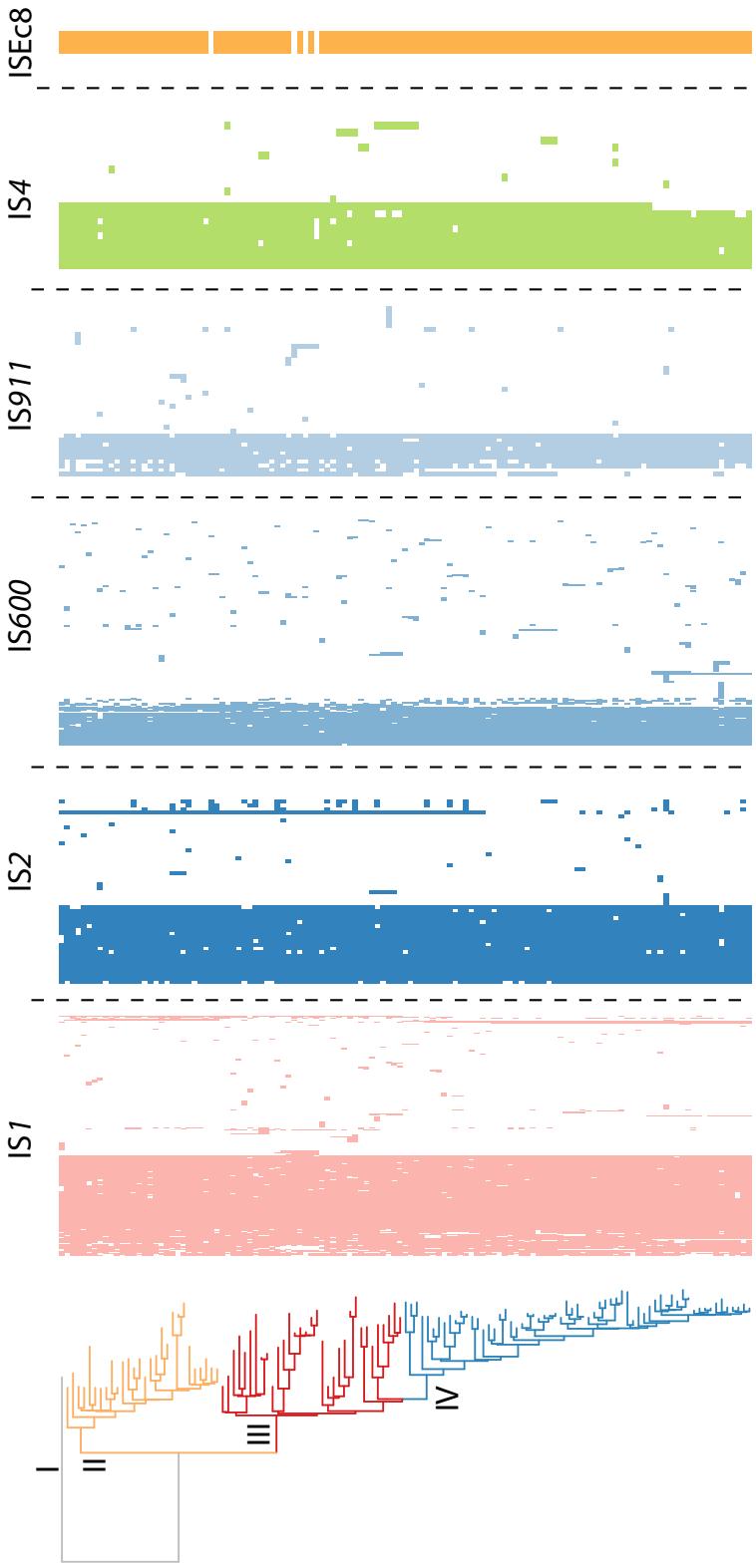

Figure 5.2: All IS insertion sites found in *S. dysenteriae* against the maximum likelihood phylogeny. Heatmaps of IS position are clustered to highly patterns between lineages. Heatmaps are divided into each IS type, and coloured shades of the same colour to indicate membership to the same IS family.

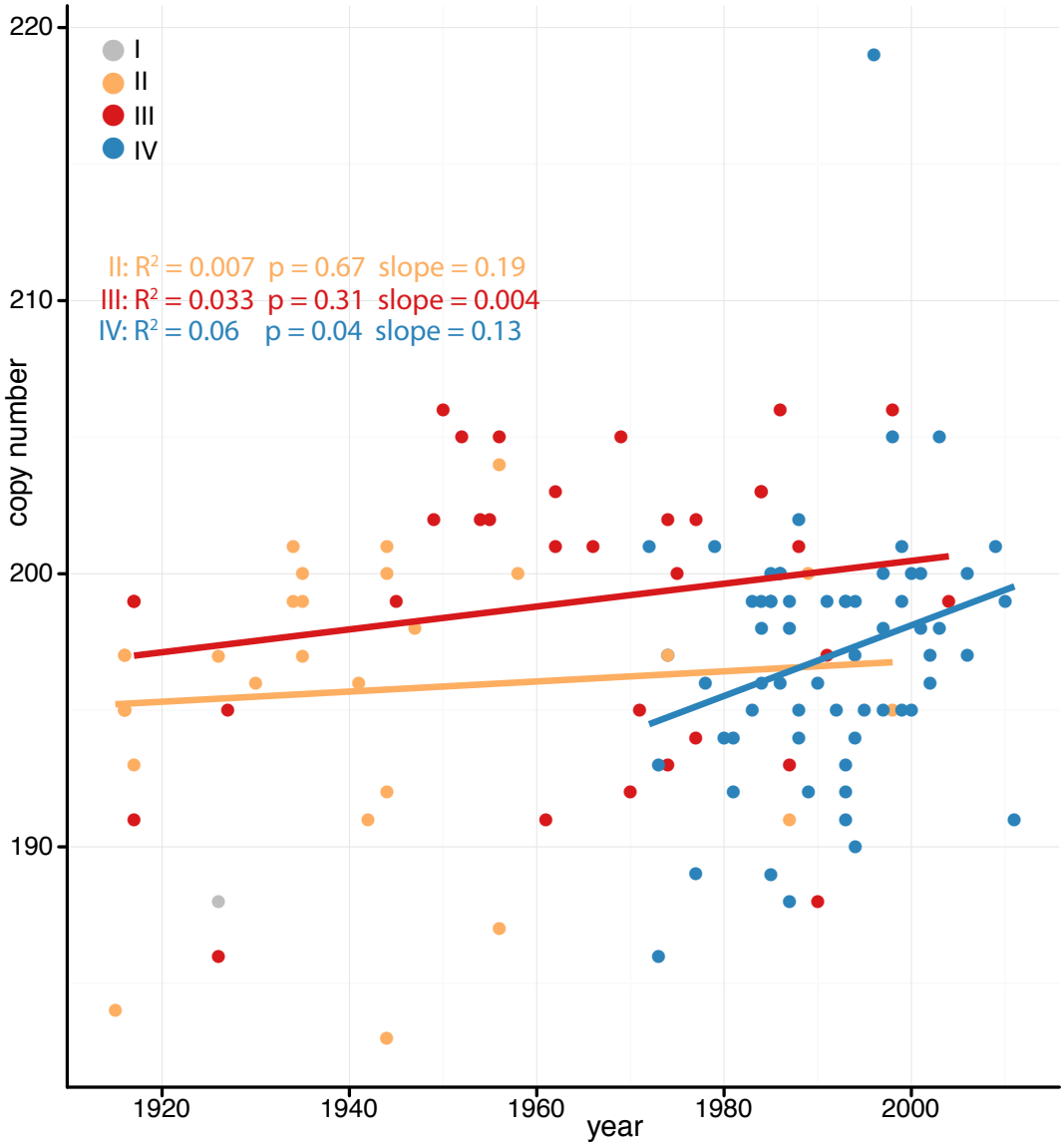

Figure 5.3: Scatterplot of total IS copy number in each genome, against the year genome was isolated. Dots coloured by lineage, as per legend. Lines show linear regressions for each lineage.

5.3.2 IS distribution in *S. flexneri*

Twelve IS were detected in *S. flexneri*, belonging to six IS families (Table 5.3). In total, 1,778 different IS insertion sites were found across all 351 *S. flexneri* genomes. Copy number varied widely by lineage, ranging from a median IS copy number of 139 in lineage 4, to a median of 224 in lineage 3 (Figure 5.4d-f). Lineages did not significantly differ in their genome size (p=0.4, Kruskal-Wallis test). In addition to copy number variation, each lineage had a distinct profile of IS insertion sites (Figure 5.6). Each lineage could be distinguished by PCA analysis of their IS insertion sites (PC1 - 36% variation, PC2 - 13% variation)(Figure 5.4c).

IS*1* contributed the most to overall IS copy number, with 40% of all IS in a genome belonging to IS*1* (Figure 5.4a, Table 5.3). The next most abundant were IS*600*, IS*2*, and IS*4*, with relative abundances of 15%, 12% and 9% respectively (Figure 5.4a, Table 5.3).

There was some variation in copy number within lineages, however, only lineage 3 showed a significant relationship between isolation date and IS copy number (p=0.009), with an accumulation of approximately 0.16 IS insertion sites year^{-1}, however the R^2 indicates that this relationship explains only 5% of the variation (Figure 5.5). IS activity was assessed by calculating the number of strain-specific insertions for each IS. IS*1* was found to be the most active IS, with 305 strain-specific IS insertions Figure 5.4b). IS*911* was also found to be very active, with 143 strain-specific IS insertions, followed by IS*2* and IS*600*, both with ~80 strain-specific sites (Figure 5.4b).

In most lineages, IS copy number in the mrca of each ancestor were similar to the median copy number in extant genomes, indicating that IS copy number has not significantly expanded since the divergence of each lineage (Table 5.4). The exception is lineage 7, which has almost doubled the number of IS copies in its genome compared to the number inferred to be present in its mrca (Table 5.4). Hence for the majority of lineages, there was no evidence of IS copy number increase over the sampling time of ~100 years, suggesting that each *S. flexneri* lineage is no longer undergoing IS expansion. However, the heatmap data, plus the strain-specific insertions, show that IS activity is still occurring, but is balanced by deletions, and this activity is not resulting in an increased copy number.

Table 5.3: IS detected in 351 *S. flexneri* genomes using ISMapper analysis.

IS	IS family	# sites across 351 genomes	mean proportion per genome	# strain-specific insertions
IS*1*	IS*1*	899	0.42	305
IS*911*	IS*3*	233	0.06	143
IS*2*	IS*3*	207	0.13	81
IS*600*	IS*3*	187	0.14	71
IS*4*	IS*4*	97	0.11	27
IS*1203*	IS*3*	72	0.038	24
ISSfl3	IS*66*	49	0.024	14
ISEhe3	IS*3*	11	0.018	4
ISSfl4	IS*110*	10	0.027	1
ISEc17	IS*3*	8	0.019	2
IS*609*	IS*200*/IS*605*	3	0.011	0
IS*150*	IS*3*	2	0.007	0

Table 5.4: Number of IS copies in the ancestors of each lineage in *S. flexneri*, as estimated by maximum parsimony, compared to the median number of IS copies found in extant genomes of that lineage. Divergence dates are those estimated by Connor *et. al.*

	ancestral number	median extant genomes	divergence date
lineage 1	191	194	1660
lineage 2	132	165	1544
lineage 3	215	224	1848
lineage 4	93	139	1341
lineage 5	162	169	1822
lineage 6	168	170	1530
lineage 7	86	146	NA

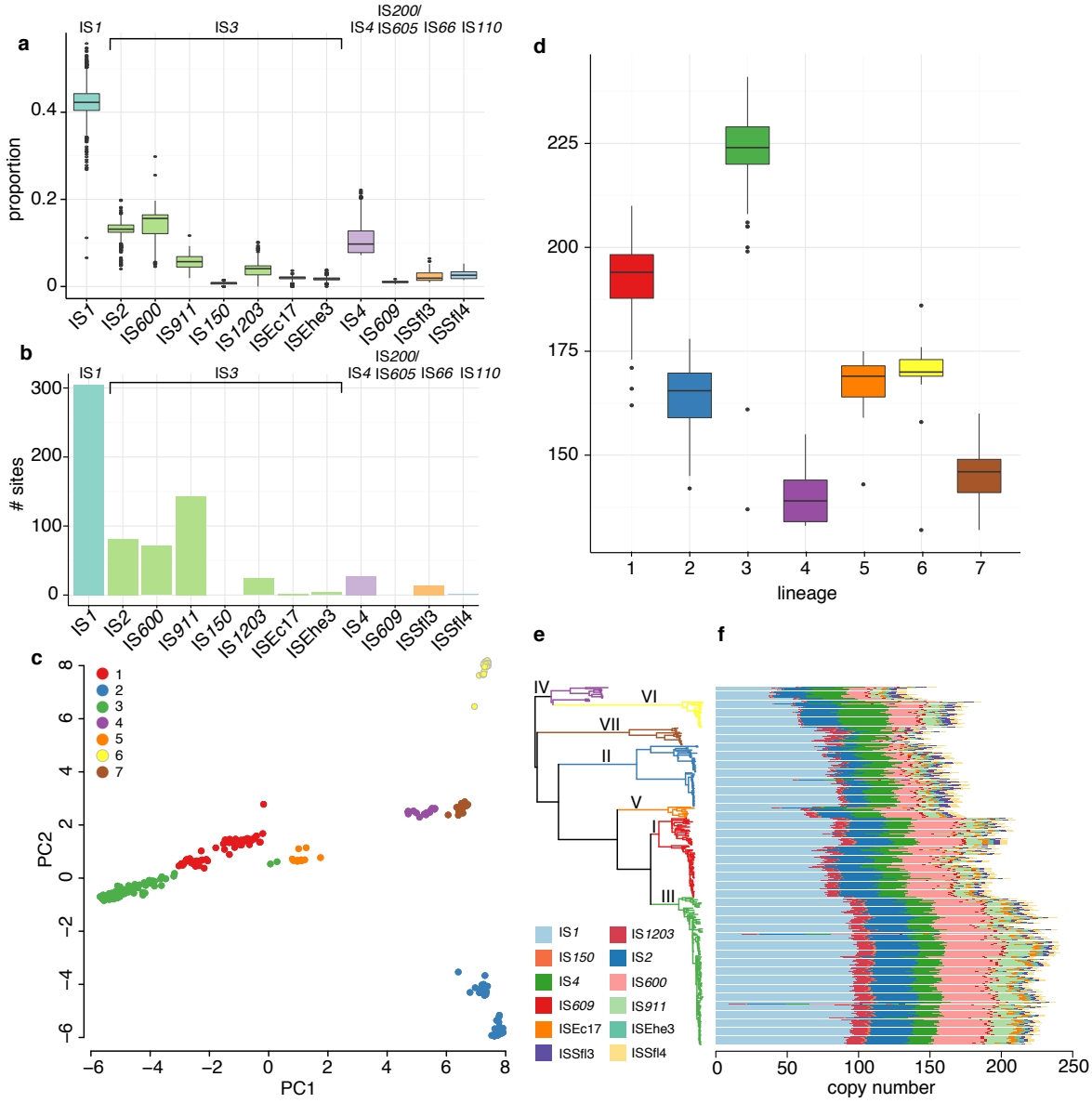

Figure 5.4: Burden of IS within *S. flexneri*. **a,** Box plots showing proportion of each IS within the genome, with IS family indicated across top. **b,** Bar plots showing the number of strain-specific insertion sites for each IS, with with IS family indicated across top. **c,** PCA of IS profiles for each genome, showing that lineages can be separated based on their IS profile. **d,** Box plots showing range of IS copy numbers within each lineage. **e,** Maximum likelihood phylogeny of all 351 *S. flexneri* genomes from Connor *et. al.*[264], with branches coloured by lineage. **f,** Bar plots showing total copy number of each IS for each genome, with bar segments coloured by IS as per legend in **e**.

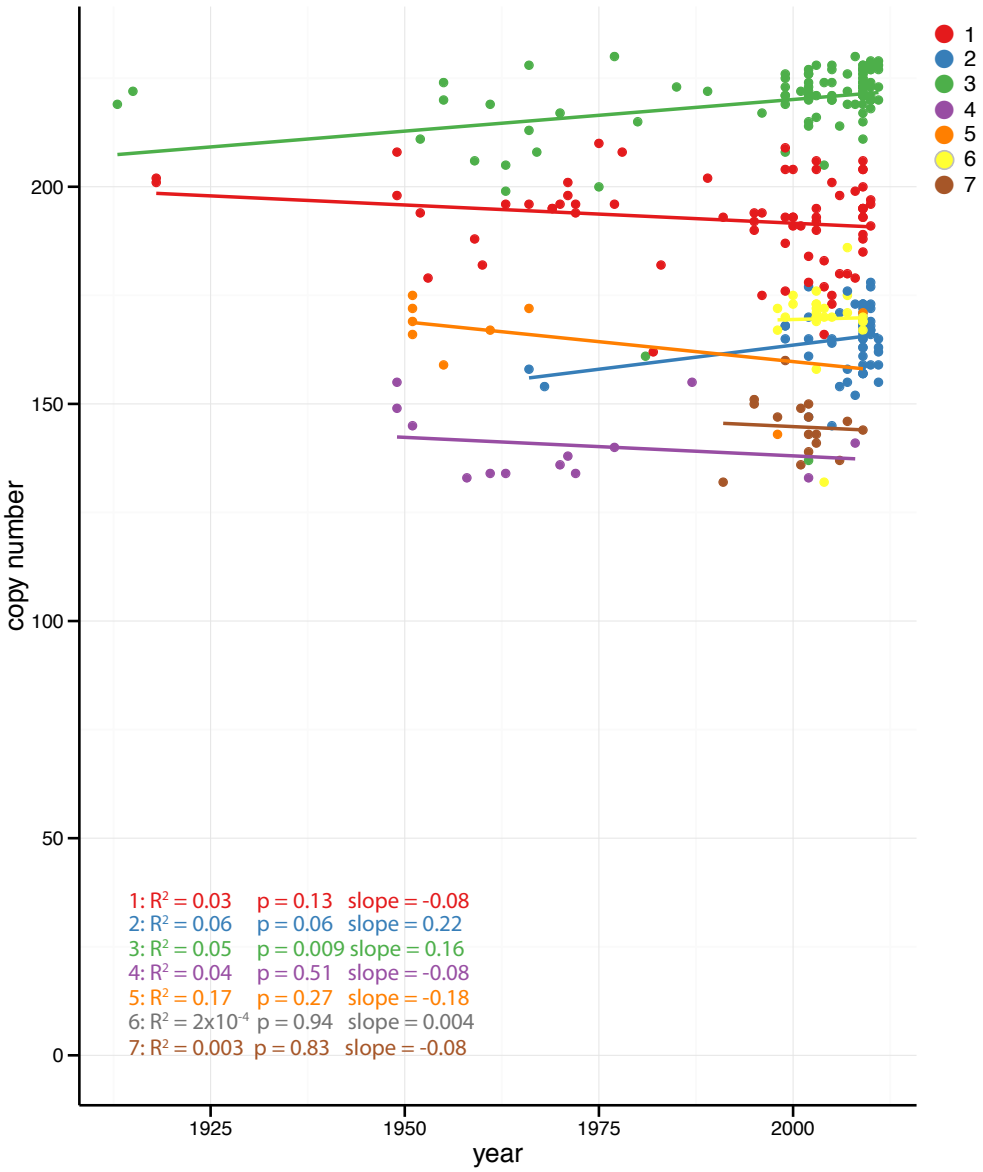

Figure 5.5: Scatterplot showing total IS copy number in each genome, against the year genome was isolated. Dots are coloured by lineage as per legend, with lines showing a best fit linear regression for each lineage.

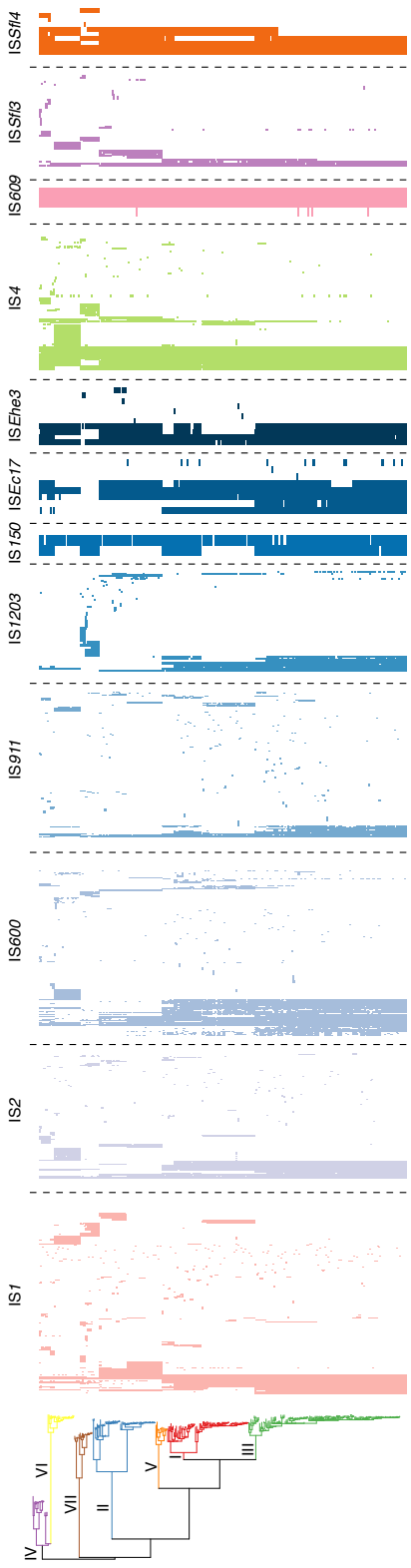

Figure 5.6: All IS insertion sites found in *S. flexneri* against the maximum likelihood phylogeny. Heatmaps of IS position are clustered to highly patterns between lineages. Heatmaps are divided into each IS type, and coloured shades of the same colour to indicate membership to the same IS family.

5.3.3 Similarities and differences between IS dynamics in each *Shigella* species

The frequency of IS insertion per base in each species varied. *S. sonnei* had the highest frequency of IS per base ($6.25x10^{-5}$, Table 5.5), followed by *S. dysenteriae* ($5.13x10^{-5}$, Table 5.5) and then *S. flexneri* ($4.52x10^{-5}$, Table 5.5). This suggests that the lower burden of IS in *S. dysenteriae* cannot be explained simply due to its small genome, as it has a higher rate of IS per base than *S. flexneri*, which has a larger genome than *S. dysenteriae*.

The three *Shigella* species had five IS in common - IS*1*, IS*2*, IS*4*, IS*600* and IS*911*. IS*1* contributed the most to overall IS copy number in all three species (42-59%, see Figure 5.7). IS*600* contributes a further 14-20%, followed by IS*2* (10-13%) and IS4 (4.5-11%) (Figure 5.7). These five IS contributed to 99% of all IS copies within *S. dysenteriae*, 86% in *S. flexneri* and 84.6% in *S. sonnei* (Figure 5.7a). *S. flexneri* contained six IS that were not found in any other *Shigella* - IS*1203*, IS*150*, ISEc17, ISEhe3, ISSfl3 and ISSfl4. *S. sonnei* also contained six IS not found in any other *Shigella* - IS*21*, IS*630*, ISEc20, ISSso1, ISSo4 and ISSo6. *S. dysenteriae* contained only a single IS that was not found in the other *Shigella*, ISEc8. Despite different specific IS types found in each *Shigella* species, these IS types belonged to only a few families (IS*3*, IS*110* and IS*66*), most of which were found in at least two *Shigella* species. The exceptions were the IS*21* family and the IS*630* family, which were only found in *S. sonnei*. For each *Shigella* species, the IS types not found in other *Shigella* species contributed to less than a quarter of overall burden in *S. sonnei* and *S. flenxeri* (Figure 5.7a).

Although there are few *S. boydii* genomes available for analysis, to investigate if the five common IS found in the other three *Shigella* species were present in *S. boydii*, I used ISSaga to identify all IS present in two *S. boydii* reference genomes, Sb227 and CDC 3083-94 (see Methods 5.2.3). All of the five common IS were also found in both *S. boydii* genomes. Overall, *S. boydii* Sb227 carried an additional 13 different IS types, in addition to the five common IS. Five of these IS (IS*1203*, ISEc20, ISSso1, ISSso6 and ISSfl3) were also found in at least one of the other three *Shigella* species. *S. boydii* CDC 3093-94 carried the same 13 IS as Sb227, and an additional four others - IS*609*, ISEc22, ISEc8 and ISKpn24. IS*609* was also found in *S. flexneri* and *S. sonnei*, and ISEc8 was also found in *S. dysenteriae*. All other IS types found in *S. boydii* were only found in *S. boydii* (IS*682*, ISEc27, ISEc47, ISSfl10, ISEc43, ISEc22, ISEc8, ISKpn14 and ISKpn24).

Each of the five common IS showed evidence of ongoing IS activity within *S. sonnei*,

S. dysenteriae and *S. flexneri*, as measured by the number of strain-specific IS insertions for each IS. To enable comparison of the number of strain-specific insertions across the three species, the numbers of strain-specific insertions were normalised by the number of strain-specific SNPs. *S. sonnei* had the highest activity level as measured by strain-specific insertions (0.12, compared to 0.04 in *S. dysenteriae* and 0.03 in *S. flexneri*, Table 5.5) In all species, IS*1* has been the most active (Table 5.5). Activity levels for the three IS*3* family members (IS*2*, IS*600* and IS*911*) differed depending on species. IS*600* was very active in *S. dysenteriae*, in comparison to *S. flexneri*, where IS*911* had a high activity level. IS*2* was most active in *S. sonnei*. In all species, IS*4* showed very low levels on activity.

Despite the evidence for expansion of IS*1*, IS*2*, IS*4*, IS*600* and IS*911* in all species, and ongoing activity of all IS, only *S. sonnei* showed evidence of ongoing net IS expansion, with increasing IS copy number in all three lineages. *S. dysenteriae* and *S. flexneri* must have undergone IS expansion at some point in their evolutionary history, however this expansion has since stabilised and IS copy number is no longer increasing.

To understand how IS content varies between pairs of strains across each *Shigella* species, I compared the percentage nucleotide diversity within and between lineages of each species with the number of IS that are either shared or not shared within and between lineages for each species (Table 5.5). All three *Shigella* species showed a high percentage of IS insertion sites that were shared between pairs of strains belonging to the same lineage (>80%, see Table 5.5). *S. dysenteriae* had a similar percentage of IS insertion sites shared between pairs of strains within lineages, as those shared between lineages (Table 5.5), indicating that *S. dysenteriae* genomes share very similar IS insertions regardless of lineage, although the lineages have diverged sufficiently to distinguish them based on IS profiles (Figure 5.1c). This relationship is similar to the within and between lineage nucloetide diversity in *S. dysenteriae* (Table 5.5). In contrast, *S. sonnei*, which had a similar level of nucleotide diversity between lineages, but had only half of IS insertion sites shared between strains of different lineages (Table 5.5). This contrast of IS content in the lineages in *S. sonnei* and *S. dysenteriae* is reflected in the number of IS insertion sites that are not shared between strains of the same or different lineages in these two species. In *S. dysenteriae*, few IS insertion sites are different between pairs of strains within the same or different lineages (Table 5.5). In *S. sonnei*, almost half of the IS insertion sites in pairs of strains of different lineages are not shared (Table 5.5). Altogether, these results show that despite similarities in nucleotide diversity between lineages in *S. dysenteriae* and

S. sonnei, IS content between lineages is more diverse in *S. sonnei* than *S. dysenteriae*.

S. flexneri has the highest level of nucleotide diversity between lineages, though strains in the same lineage have a similar nucleotide diversity level to all strains in *S. dysenteriae* (Table 5.5). Differences in IS content between *S. flexneri* lineages are greater than differences in IS content found between *S. sonnei* or *S. dysenteriae* lineages - in *S. flexneri*, only a third of IS insertion sites are shared between pairs of strains from different lineages (Table 5.5).

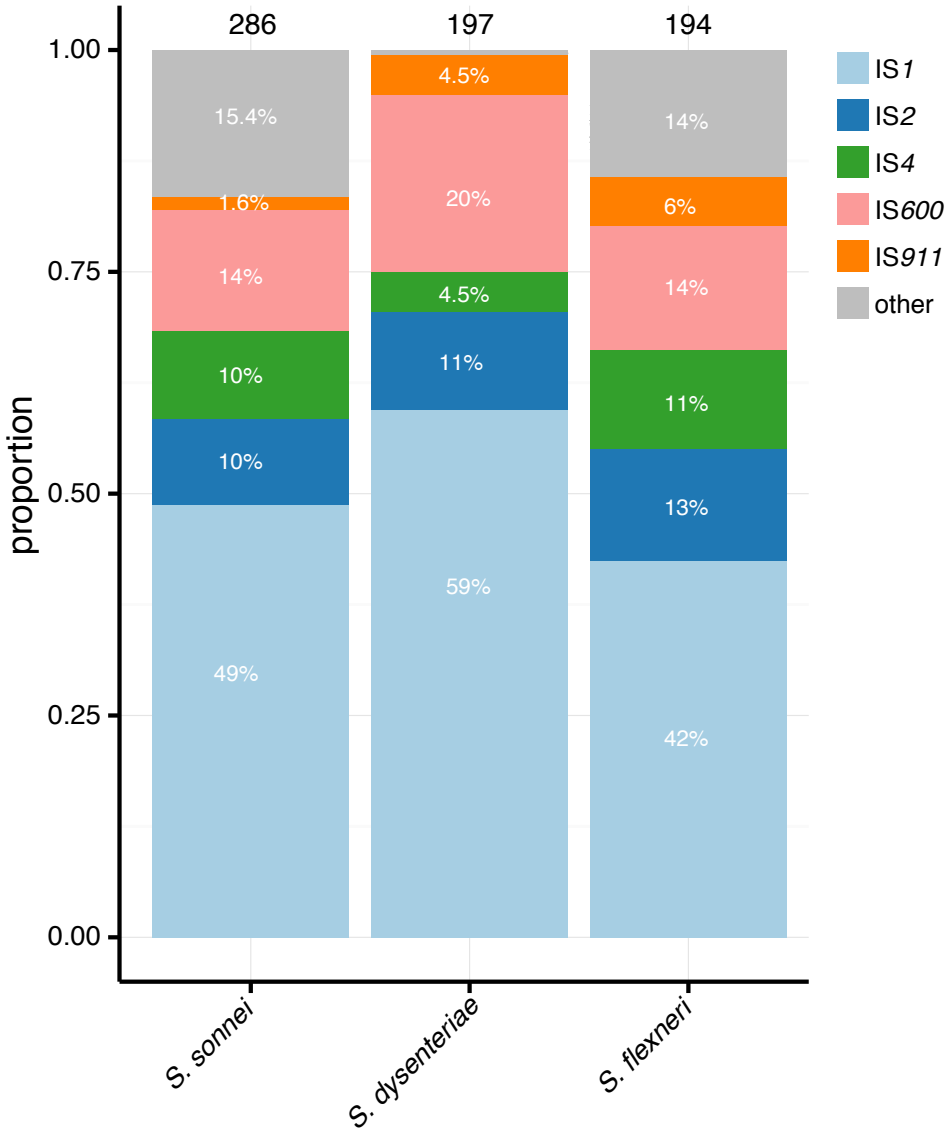

Figure 5.7: Comparison of burden for five IS found in each *Shigella* species. Proportion of five common IS, IS*1*, IS*2*, IS*4*, IS*600* and IS*911* in all three *Shigella* species. Median IS copy number in each species is labelled in black at the top of each bar.

Table 5.5: Comparison of nucleotide and IS diversity in each *Shigella* species. For *E. coli*, the mean value across all genomes is shown, with the minimum and maximum values in brackets.

		S. sonnei	*S. dysenteriae*	*S. flexneri*	*E. coli*
Reference genome size (bp)		4988504	4369232	4607202	5108280 (4686000 - 5547000)
mean % nt divergence	*within lineages*	0.004%	0.007%	0.012%	
	between lineages	0.023%	0.017%	0.089%	
median IS copy number		286	197	194	15 (4 - 44)
mean IS per base		6.25×10^{-5}	5.13×10^{-5}	4.52×10^{-5}	
median IS*1* copy number		141	117	85	
proportion of five common IS		84.6%	86%	99%	
Pairwise shared IS	*within lineages*	265 (84%)	184 (87%)	187 (81%)	
	between lineages	195 (53%)	178 (82%)	91 (33%)	
Pairwise non-shared IS	*within lineages*	51 (16%)	27 (13%)	44 (19%)	
	between lineages	174 (47%)	39 (18%)	195 (67%)	
strain-specific IS		0.12 (593)	0.04 (303)	0.03 (672)	
strain-specific IS*1*		0.04 (220)	0.02 (112)	0.014 (305)	
strain-specific IS*2*		0.03 (128)	0.003 (21)	0.004 (81)	
strain-specific IS*4*		0.004 (21)	0.0008 (6)	0.001 (27)	
strain-specific IS*600*		0.01 (74)	0.02 (137)	0.003 (71)	
strain-specific IS*911*		0.01 (62)	0.004 (27)	0.007 (143)	
species-specific IS		IS*21*, IS*630*, ISEc20, ISSso1, ISSo4,	ISEc8	IS*1203*, IS*150*, ISEc17, ISEhe3, ISSfl3, ISSfl4	

5.3.4 Comparing IS in *E. coli* with *Shigella*

Previous studies have shown that *Shigella* species have reduced genomes and a large number of IS, compared to *E. coli*. The previous section demonstrates that *S. sonnei*, *S. dysenteriae* and *S. flexneri* have each undergone expansions of the same five IS. *S. sonnei* has the largest genome of the three (4.9 Mbp), similar to most other *E. coli*, and is still accumulating IS, while *S. flexneri* and *S. dysenteriae* have reduced genomes (4.6 Mbp and 4.3 Mbp, respectively) and each appear to have reached IS equilibrium. IS transposition is still occurring in *S. dysenteriae* and *S. flexneri*, but this is balanced by loss of IS, so their overall IS copy numbers are stable. This section aims to place these findings in the context of IS dynamics in the wider *E. coli* population, and answer the following questions:

i) are the five IS found that have undergone expansion in all *Shigella* species also found in other *E. coli*? and;

ii) are other pathogenic lineages of *E. coli* undergoing IS expansion, or is this a unique feature of *Shigella*?

To investigate burden of the five common IS within *E. coli*, 1000 publically available genomes from the GenomeTrackr project, which involves sharing foodborne bacteria by public health agencies, were investigated. Because of the wide range and diversity of the *E. coli* genomes in this dataset, which share only a small component of their core genome, a reference-based approach such as ISMapper analysis was not suitable for this comparison. An alternative mapping-based approach was therefore applied to estimate copy number in the *E. coli* and *Shigella* genomes (see Methods section 5.2.6). The IS copy number in each *Shigella* species estimated by this method was different to the IS copy number estimated using ISMapper (medians 311 vs 292, respectively, in *S. sonnei*; 147 vs 198 in *S. dysenteriae*; 236 vs 198 in *S. flexneri*). The IS copy number estimated using a reference-free approach was usually inflated compared to the ISMapper estimate.

Using the reference-free method to estimate IS depth ratio, across all *E. coli* the median burden of these five IS was 15 (range 4 to 44), significantly lower than the IS copy number found in each *Shigella* species (medians: *S. sonnei*, 292; *S. flexneri*, 198; *S. dysenteriae*, 198). These data showed no evidence of expansion of the five IS common to *Shigella* in any of the *E. coli* groups (Figure 5.8).

As the five common IS were not found to be expanded in the pathogenic lineages of the 1000

E. coli genomes, I asked whether pathogenic lineages of *E. coli* may contain high burdens of IS that are not one of the five common IS found in *Shigella*. To explore this, expanded datasets of three pathogenic *E. coli* lineages were collected: 82 ST131 (UPEC) genomes, 199 ST11 (EHEC) genomes and 36 O104:H4 outbreak genomes. The full complement of IS within these lineages was investigated using ISMapper (Table 5.6), to map IS insertions relative to a reference genome within each lineage, and the results compared with the ISMapper results for *Shigella* (Figure 5.9a). There were a median of nine IS insertions found in the ST131 genomes, with a maximum of 24 IS insertions, significantly lower than the copy numbers found in *Shigella* using ISMapper analysis (Figure 5.9a). The median IS copy number found in ST11 was 49 (Figure 5.9a). Median IS copy number in O104:H4 was 17 (Figure 5.9a). These data showed that all three pathogenic lineages had significantly fewer IS than the three *Shigella* species ($p=2.2x10^{-16}$ for all comparisons, Wilcox test). Additionally, unlike each *Shigella* species, none of these pathogenic lineages had a high IS*1* burden (Table 5.6). In all cases, a different IS contributed the most to burden - ISEc12 in ST131 (copy number 0-6), IS*1203* in ST11 (copy number 4-35), and ISEc23 in O104:H4 (copy number 1-12) (Figure 5.9b).

The previous results show that overall, the five common IS in *Shigella* are present but not expanded in the wider *E. coli* population (Figure 5.8), and that the most well known pathogenic lineages of *E. coli* investigated do not have a high IS copy number compared to that of *Shigella* species (Figure 5.9a). The most striking difference in IS copy number between *Shigella* and other *E. coli* is the prevalence of IS*1* within *Shigella* genomes, although all five IS showed evidence of expansion in *Shigella* (Figure 5.8). As many of the *Shigella*-expanded IS, including IS*1*, are present in most *E. coli* genomes, this raises the question of why they have not undergone expansion in other *E. coli* lineages? Increased IS*1* transposition activity within *Shigella* could be due to multiple factors, such as: mutations in the frameshift region altering proportions of InsA and InsAB' product which control the rate of transposition, changes in host factors that influence transposition activity, or relaxation of purifying selection in *Shigella* genomes following host restriction.

To understand why IS*1* is expanded in all three *Shigella* species, and none of the *E. coli* lineages, I compared IS*1* sequences extracted from the three *Shigella* species and ten *E. coli* reference genomes. The resulting phylogeny shows that the majority of IS*1* sequences present each *Shigella* species are derived from sequences present in the wider *E. coli* population, which have then diversified within each *Shigella* population (Figure 5.10). There has been a small amount

of transfer of IS*1* sequences from *S. dysenteriae* into *E. coli* APEC01 (Figure 5.10). No mutations were observed that could explain the expansion of IS*1* within *Shigella*. When comparing alignments of IS*1* sequences across each *Shigella* species, none of the IS*1* sequences in each *Shigella* species had the A_7C or GA_2A_3C motif within the frameshift region of *insA* and *insB*, which has been shown to increase transposition of IS*1*[374,375].

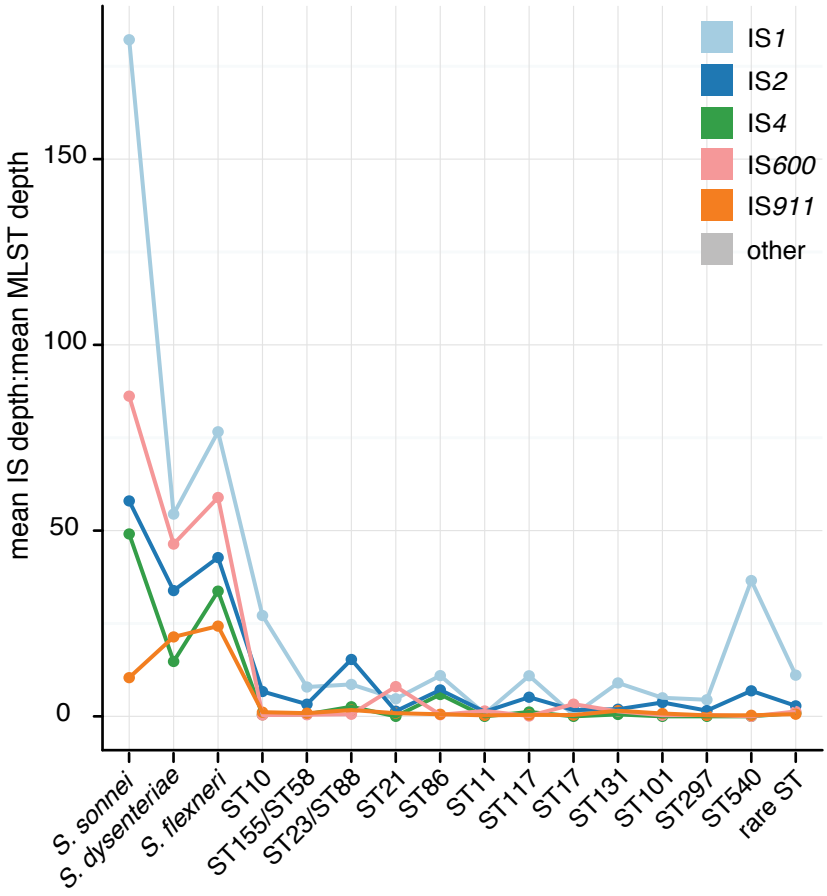

Figure 5.8: Burden and proportions of the five common IS in *Shigella* and *E. coli* Line graph indicating estimated burden of each of the five common IS in all three *Shigella* species and several *E. coli* STs. Burden is estimated as the ratio of mean read depth of IS against mean read depth across MLST genes.

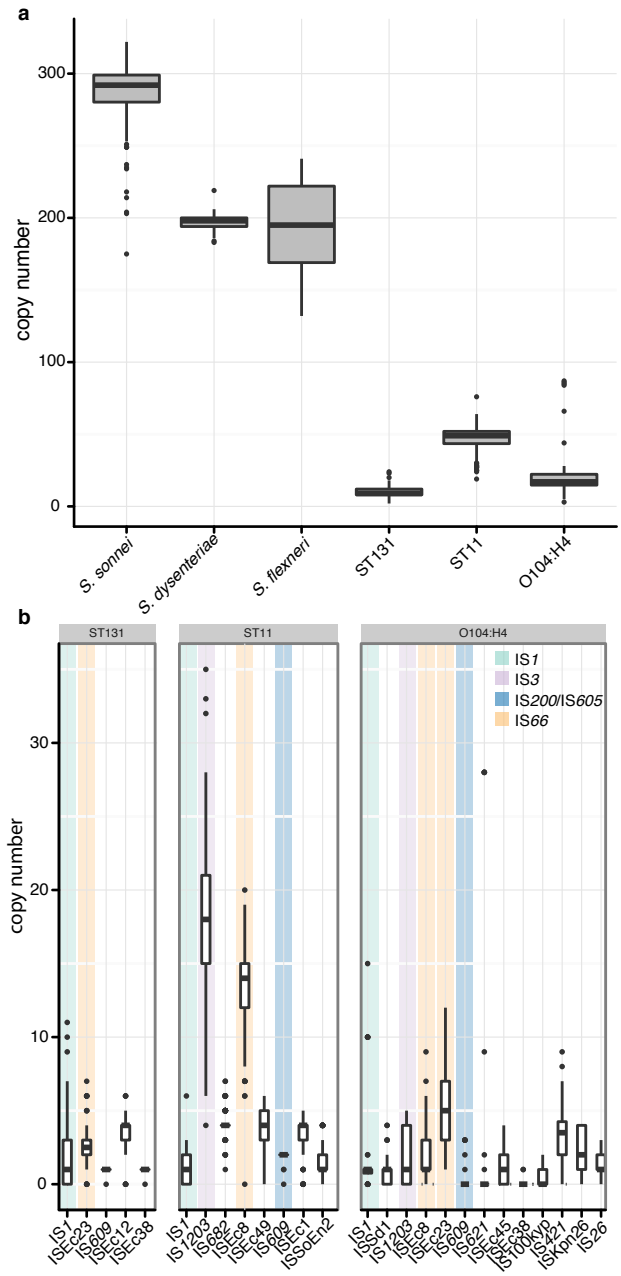

Figure 5.9: Comparison of IS content in the three *Shigella* species and three pathogenic *E. coli* lineages. a, Box plots showing IS copy number in each *Shigella* species and each pathogenic lineage of *E. coli*, estimated using ISMapper. **b,** IS copy numbers for each pathogenic *E. coli* lineage, estimated using ISMapper. Boxplots illustrating IS copy number for all IS detected in ST131, ST11 and O104:H4. Highlighting indicates IS found in at least one *Shigella* species, coloured by IS family, as per legend.

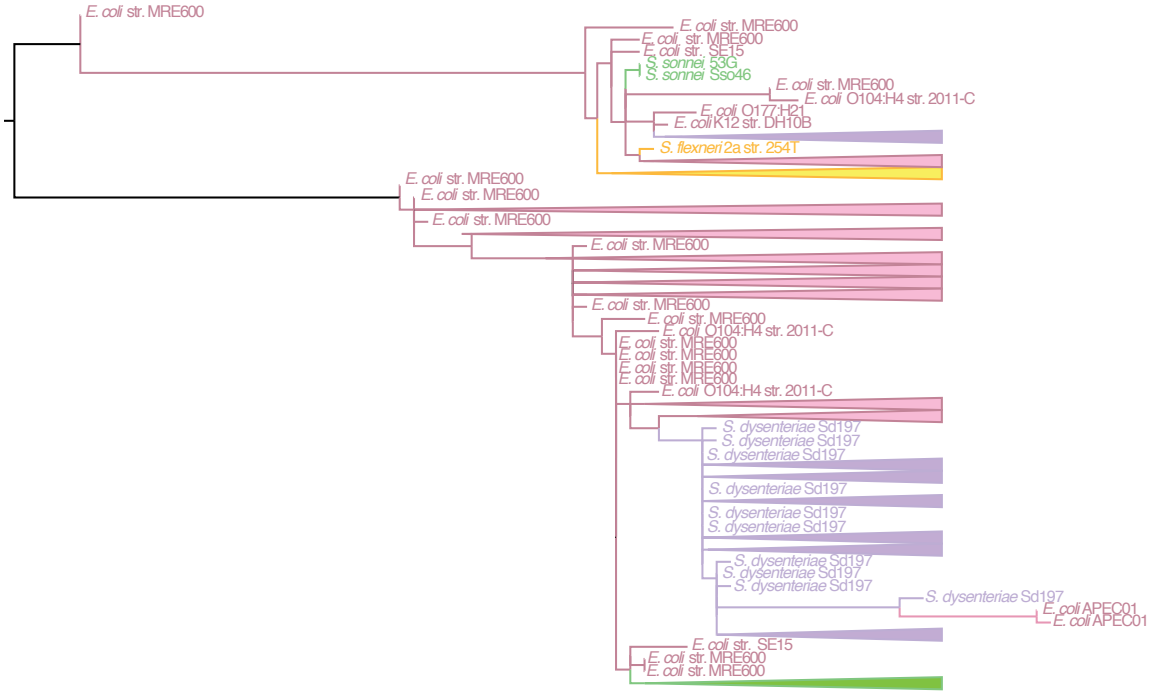

Figure 5.10: Maximum likelihood phylogeny of 827 IS*1* sequences sourced from ten *E. coli* reference genomes and five *Shigella* genomes. The phylogeny is midpoint rooted. Branches have been collapsed to highlight relationships between groups. Branches are coloured by the species the IS*1* sequence was extracted from. Pink - *E. coli*; green - *S. sonnei*; purple - *S. dysenteriae*; orange - *S. flexneri*.

Table 5.6: IS found in each pathogenic *E. coli* lineage, with their family and the mean proportion per genome.

Lineage	IS	Family	# sites	mean proportion per genome	# strain-specific insertions
ST131	IS*1*	IS*1*	127	0.17	97
	ISEc23	IS*66*	84	0.25	57
	ISEc12	IS*21*	35	0.36	15
	ISEc38	ISL3	1	0.11	0
	IS*609*	IS*200*/IS*605*	1	0.11	0
ST11	IS*1203*	IS*3*	625	0.38	
	ISEc8	IS*66*	96	0.29	30
	ISSoEn2	IS*256*	20	0.03	3
	IS*682*	IS*66*	13	0.09	1
	ISEc49	IS*66*	7	0.07	1
	IS*1*	IS*1*	7	0.02	5
	IS*609*	IS*200*/IS*605*	6	0.04	0
	ISEc1	ISAs1	6	0.07	0
O104:H4	ISEc23	IS*66*	63	0.26	19
	IS*1203*	IS*3*	59	0.07	22
	IS*421*	IS*4*	58	0.15	9
	ISEc8	IS*66*	49	0.11	13
	IS*621*	IS*110*	39	0.04	1
	IS*1*	IS*1*	29	0.07	17
	IS*26*	IS*26*	19	0.07	0
	ISKpn26	IS*5*	18	0.12	2
	ISEc45	IS*110*	17	0.05	4
	ISSd1	IS*3*	15	0.04	4
	IS100kyp	IS*21*	8	0.02	0
	IS*609*	IS*200*/IS*605*	5	0.007	1
	ISEc38	ISL3	5	0.001	2

5.3.5 Functional impact of IS on gene inactivation in *Shigella*

As discussed in Chapter 1, each species of *Shigella* has independently evolved to cause a similar disease phenotype. Much of that evolution has involved pseudogene formation, which results in loss of gene function, and IS have played a significant role in this. In this section I start by investigating gene inactivation within *S. dysenteriae* and *S. flexneri* individually, before comparing gene inactivation across all three *Shigella* species. The following questions are addressed:

i) do all three *Shigella* species have similar levels of purifying selection;

ii) are there gene functions significantly enriched for inactivating mutations; and

iii) are there genes or functional groups that are frequently inactivated in two more more *Shigella* species?

5.3.5.1 Gene inactivation in *S. dysenteriae*

To investigate gene inactivation within *S. dysenteriae*, all non-synonymous SNPs and intragenic indels were collated. All lineages had similar numbers of inactivated genes (median 289). One quarter (24.6%) of all *S. dysenteriae* genes were inactivated in one or more genomes, and 9.3% of genes were inactivated by IS in one or more genomes. Double this number of genes were inactivated by mutation (18.9%) in one more more genomes, with 3.5% of genes inactivated by both IS and mutation. IS insertions were found in 0.011% of genic bases, compared to 0.037% of intergenic bases (ratio 0.27, $p < 2.2x10^{-16}$, Table 5.7), demonstrating that insertions within coding regions are under purifying selection.

To investigate the effect of gene inactivation in *S. dysenteriae*, each gene was assigned to a biological system using RAST. Only 36% of genes were assigned a biological system. IS-mediated inactivation was most common amongst genes of unknown function (85% of genes not assigned to a system vs 65% of genes assigned to a system). Genes involved in maltose and maltodextrin utilisation (p=0.032, OR 6.89, 95% CI [1.19 - 27.38]), phage packaging machinery (p=0.027, OR 24.76, 95% CI [1.76 - 347]) and unknown carbohydrate metabolism (p=0.027, OR 7.61, 95% CI [1.30 - 30.95]) were enriched for IS-mediated gene interruption.

One hotspot of gene inactivation (>3 IS insertions in a 1000 bp window) were the invasion plasmid antigen genes located on the chromosome. Two of the six invasion plasmid antigen genes (*ipaH_3* and *ipaH_6*) in *S. dysenteriae* were interrupted by three or more independent IS insertion sites. Previous work in *S. flexneri* has shown that chromosomal *ipaH* genes are secreted via the T3SS, however deletion mutants show no changes to pathogenesis in mouse models infected with *S. flexneri*, suggesting that the chromosomal *ipaH* genes may have redundant functionality[376].

5.3.5.2 Gene inactivation in *S. flexneri*

Across the entire *S. flexneri* data set, there were 1,741 (44.6%) genes inactivated in at least one genome, suggesting that that these genes are likely not essential for growth. Of all genes inactivated in *S. flexneri*, 18.3% were inactivated by IS and 37.6% by mutational interruptions. Overall, the number of genes inactivated in *S. flexneri* is twice as many as the number of inactivated genes found in either *S. sonnei* or *S. dysenteriae*. This is likely because the sample set for *S. flexneri* represents a much longer amount of evolutionary time. Within specific lineages, the percentage of inactivated genes is much closer to the proportions found in *S. sonnei* or *S. dysenteriae*. Lineages 1 and 3 have similar proportions of inactivated genes (17.6% and 20% respectively), both of which are similar in proportion to the amount of inactivation found in *S. sonnei* and *S. dysenteriae*. Both of these lineages have been diversifying for < 300 years, similar to the evolutionary timescale of *S. sonnei* and *S. dysenteriae*. The rate of IS insertion per base was 0.03% in genic regions, versus 0.11% of bases in intergenic regions (ratio 0.26, $p < 2.2x10^{-16}$, Table 5.7), indicating that purifying selection is occurring within the coding regions of *S. flexneri*.

Only 13 gene interruptions were conserved amongst all *S. flexneri* genomes. Nine of these 13 genes were hypothetical. One of the 13 genes (SF3448, a sn-glycerol-3-phosphate dehydrogenase) was interrupted by IS*1*. The remaining gene interruptions conserved across all genomes were *mukB* (SF0920), a cell division protein; *setB* (SF2255), a sugar transport protein[377]; and *ravA* (SF3826), a two component regulator. In *E. coli*, *ravA* has been shown to interact with *cadA*, in the cadaverine production pathway[378]. As *cadA* has been inactivated in all *Shigella* species[233], inactivation of *ravA* likely represents degradation of an already inactive pathway. The *mukB* gene has been shown to involved in the division of chromosomes during cell replication in *E. coli*, and strains with this gene inactivated are unable to form colonies[379].

However, mutations in several other genes have been shown to allow strains to compensate for the inactivation of *mukB* in *E. coli*[380–384].

Gene enrichment analysis assigned 52% of genes to a biological system. IS-mediated gene interruptions were common in genes of unknown function (71% of genes not assigned to a system vs 48% of genes assigned to a system). Genes involved in di- and oligo-saccharide synthesis ($p=2.7x10^{-7}$, OR 3.6, 95% CI [1.94 - 6.33]), type VII protein secretion systems ($p=2.7x10^{-6}$, OR 7.24, 95% CI [3.04 - 16.84]), and motility and chemotaxis (p=0.029, OR 4.88, 95% CI [1.27 - 16.4]) were significantly enriched for IS interruption across all *S. flexneri* genomes. When including genes that were also inactivated by mutation, in addition to the previous pathways, metabolism of central aromatic intermediates ($p=4.34x10^{-6}$, OR 12.49 95% CI [2.89 - 112.97]) was also enriched.

5.3.5.3 Comparison of gene inactivations between *Shigella* species

All three *Shigella* species were found to be undergoing purifying selection for IS insertion sites within the coding regions of their genomes. The ratio of genic to intragenic IS insertions was < 1 in all three species - 0.17 (*S. sonnei*), 0.27 (*S. dysenteriae*) and 0.26 (*S. flexneri*), indicating that purifying selection of IS insertion sites is ocurring in all three species. However, this purifying selection appears to be stronger in *S. sonnei* than the other two species. *S. dysenteriae* had the fewest IS insertions in coding regions (0.011% bases), compared to 0.016% in *S. sonnei* and 0.03% in *S. flexneri* (Table 5.7). All three species had more genes inactivated by mutation than IS (Figure 5.11).

S. dysenteriae is the smallest of the *Shigella* genomes, and has undergone the most genome decay. To investigate whether the populations of *S. flexneri* and *S. sonnei* are on a similar evolutionary path to *S. dysenteriae*, I compared inactivated genes in *S. flexneri* or *S. sonnei* to inactivated genes in *S. dysenteriae*. Out of the 868 inactivated genes in *S. sonnei*, 91 of these were also inactive in *S. dysenteriae*, and 643 were completely absent from *S. dysenteriae* (i.e. there were no orthologs present, and so assumed to have already succumbed to genome decay). Overall, 84% of genes inactive in *S. sonnei* were either also inactive in *S. dysenteriae* (91 genes) or completely absent (i.e. there were no orthologs present, and so assumed to have already succumbed to genome decay, 643 genes) (Figure 5.12). From the 1,741 genes inactive in *S. flexneri*, 79.7% of these genes were either inactive in *S. dysenteriae* (169 genes) or absent entirely (1209 genes) from

S. dysenteriae (Figure 5.12). There were 288 genes that were inactive in both *S. flexneri* and *S. sonnei*, and 74% were either inactive or absent from *S. dysenteriae* (37 genes and 176 genes, respectively, Figure 5.12).

Table 5.7: Comparison of IS insertion rates and pairwise shared and non-shared pseudogenes for each *Shigella* species.

		S. sonnei	*S. dysenteriae*	*S. flexneri*
Genic insertion rate		0.016%	0.011%	0.03%
Intergenic insertion rate		0.083%	0.037%	0.11%
Pairwise shared pseudogenes	*within lineages*	131 (76%)	177 (78%)	173 (75%)
	between lineages	74 (37%)	165 (71%)	122 (33%)
Pairwise non-shared pseudogenes	*within lineages*	41 (24%)	48 (22%)	62 (25%)
	between lineages	127 (63%)	68 (29%)	265 (67%)

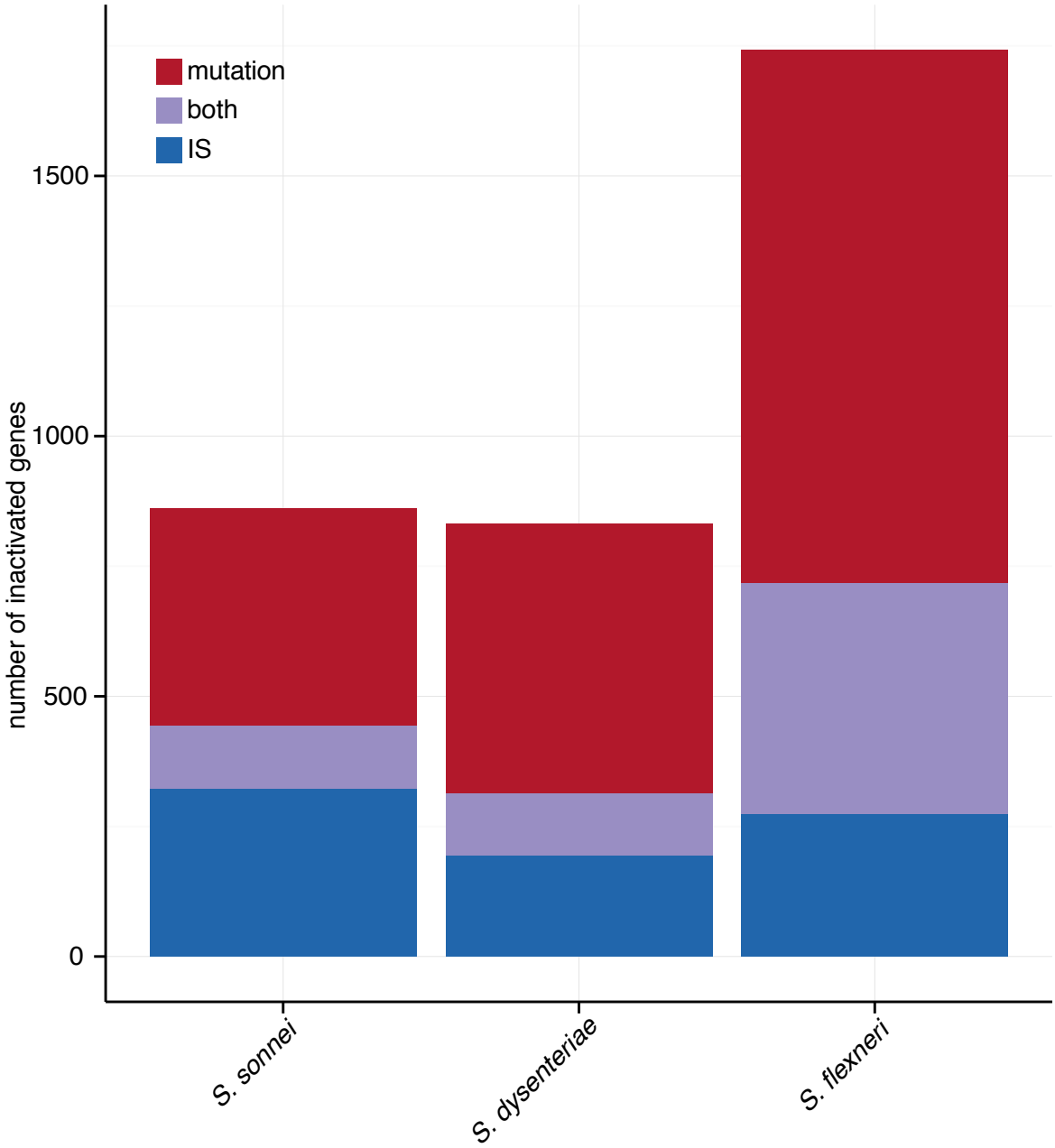

Figure 5.11: Number of genes inactivated in each *Shigella* species. Colours inside bar plots indicate the cause of gene inactivation, either IS interruption, mutation, or both.

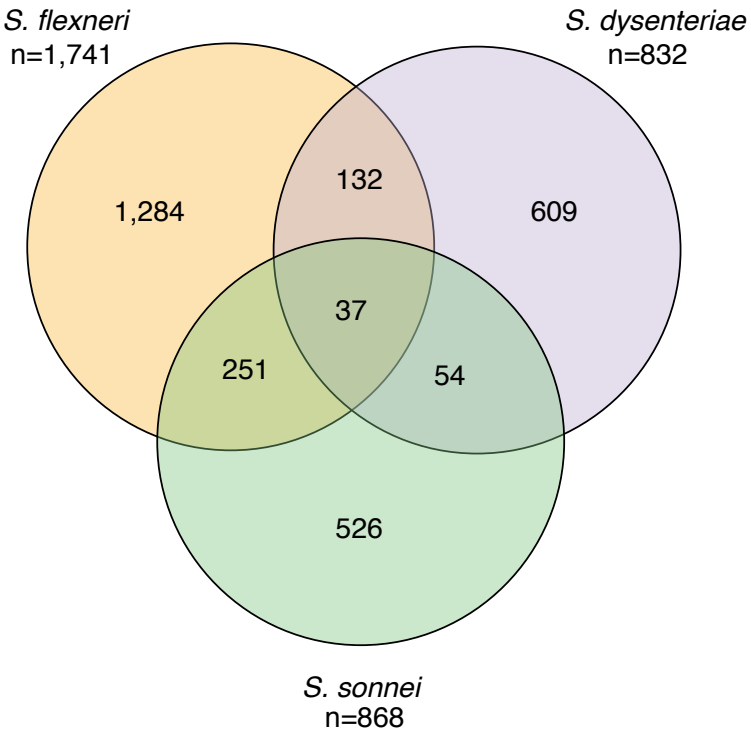

Figure 5.12: Comparison of number of genes inactivated in each *Shigella* species.

5.4 Discussion

5.4.1 Strengths and limitations

This chapter builds on the framework established in Chapter 4 to explore the dynamics of IS in two additional species of *Shigella*, *S. dysenteriae* and *S. flexneri*, and briefly explores the IS found in two *S. boydii* reference genomes. By investigating these additional *Shigella* species, the results in Chapter 4 can be placed into the wider context of *Shigella* evolution. This chapter has provided novel insights into the behaviour of IS within *Shigella*, and the similarities and differences of IS dynamics between species. By combining the population structures of *S. dysenteriae* and *S. flexneri* with IS data obtained using ISMapper, this chapter was able to:

i) show that IS copy number in *S. dysenteriae* is similar between lineages, and overall lower than the IS copy number found in *S. sonnei*;

ii) show that IS copy number in *S. flexneri* varies amongst lineages, but overall burden is lower than IS copy number in *S. sonnei*;

iii) demonstrate that five IS (IS*1*, IS*2*, IS*4*, IS*600* and IS*911*) were found to be common to all three *Shigella* species, and these IS were also present in *S. boydii*;

iv) that the five IS found in all three *Shigella* species are also present in the wider *E. coli* population, but have not undergone signficant expansion in the wider *E. coli* population, or in the well-known pathogenic lineages of *E. coli*; and

v) that on-going gene decay is likely occurring in *S. sonnei* and *S. flexneri*, but *S. dysenteriae* has largely stabilised its genome content.

However, this chapter had some limitations surrounding the availability of data for both *S. dysenteriae* and *S. flexneri*. Only one serotype of *S. dysenteriae* was examined, *S. dysenteriae* Sd1. From the fifteen serotypes of *S. dysenteriae*, Sd1 is responsible for all of the major *S. dysenteriae* outbreaks since the late 19th century, and is the most studied and well sequenced *S. dysenteriae* serotype.

Several serotypes of *S. flexneri* were included in the dataset from Connor *et. al.*[264] examined in this chapter, with the exception of serotype 6, as this serotype clusters with mostly *S. boydii* strains[263]. Despite having more *S. flexneri* genomes than either *S. sonnei* or *S. dysenteriae*, the

majority of these genomes belonged to only two lineages, 1 or 3. Additional data from the under-represented lineages, 4, 5, 6 and 7, would help increase our understanding of IS lineage differences in *S. flexneri*. Comparisons between *S. flexneri* and the other two *Shigella* species were also complicated by the fact that *S. flexneri* is much older and more diversified than the other two *Shigella* species. It would be more meaningful to compare each individual *S. flexneri* lineage with *S. sonnei* and *S. dysenteriae*, but more sampling of all *S. flexneri* lineages is required to achieve this.

S. boydii was not investigated in detail in this chapter, due to the limited number of publically available genomes. Some work has been done on the population structure of *S. boydii*, and this has revealed that *S. boydii* genomes are found throughout all three of the main *Shigella* clades, complicating the analysis of this species[361]. However, inspection of the two complete *S. boydii* genomes Sb227 and CDC 3083-94, revealed that these two strains carry the five IS common to the other three *Shigella* species examined in this thesis. In addition, the two *S. boydii* genomes carried several other IS types, some of which were found in at least one of the other three *Shigella* species, and some of which were found only in *S. boydii*. Additional *S. boydii* data would provide an interesting comparison of IS variability in this species compared to the other three species, however the three species analysed in detail in this study represent the vast majority of the global burden of dysentery[202,385].

Important limitations were exposed in this framework when IS were compared between each *Shigella* species and *E. coli*. As *E. coli* is a large and diverse group, the reference-based approach relied on by ISMapper was not suitable for investigating IS amongst the entire group, as a close reference is required to accurately detect IS. As no such reference exists for all *E. coli*, a reference-free approach, such as an assembly based IS detection method may be more suitable. However, assembling all genomes would be computationally intensive, and copies of IS inserted within other IS would still not be detected. ISMapper was shown to underestimate IS copy number in *Shigella* compared to using a IS depth ratio method, however it is unclear which method is more accurate. In two of the three *Shigella* species, the reference-free depth approach detected more IS copies than ISMapper. There are two main reasons for this - firstly, there may be regions in the genome, that are not present in the reference, where IS are located that ISMapper will not be able to detect. There may have been additional IS present on the virulence plasmid or other plasmids in some genomes that could not be detected using a chromosomal reference with ISMapper. Secondly, IS can insert within other IS - ISMapper will only be able to detect

the outermost IS. Additional IS copies found in the depth method may have come from this type of situation in some genomes. However, the IS copy number estimated using the IS depth ratio method assumes that there are only single copies of each MLST loci. If any of the MLST loci have duplicated within the genome, this IS depth ratio will not be an accurate reflection of IS copy number.

As discussed in Chapter 4, the virulence plasmid was not included in this analysis, as there were few genomes containing plasmid sequence at appropriate levels of coverage and depth for analysis with ISMapper. If plasmid sequence was available for all genomes, future investigation could examine if differences in IS copies per base were similar between chromosomal and plasmid sequence for each species. IS types found on plasmid sequences in each *Shigella* species could be compared. Additional unanswered questions include: are there any common IS found amongst all plasmid sequences across all *Shigella* species? As IS*1* is found on the virulence plasmid[341], what are the dynamics of IS*1* on the plasmid, and how does this compare across *Shigella* species?

Finally, this chapter has not fully explored the biological implications of gene inactivation in each *Shigella* species. Understanding the pathways genes belong to and the functional significance of changes in their expression is a complex area of research. The RAST annotation tool assigns each identified gene to FIGfams, which can then be matched to a hierarchical biological system within the SEED framework[344]. However, even with biological system annotation using RAST, it can still be difficult to understand relationships between genes and the functional impact of gene loss without additional experimental work. The researchers behind RAST are attempting to address this problem through the construction of metabolic models for bacterial species[386], however this area of research is still in its infancy. In each *Shigella* species, at least a third of genes could not be assigned to a system, and had an unknown function, so a significant amount of experimental work still needs to be undertaken to address these questions in more detail.

5.4.2 Contribution of IS to the evolution of each *Shigella* species

As discussed in section 5.1, each *Shigella* species has a smaller genome than *E. coli*, with *S. dysenteriae* having the smallest genome and *S. sonnei* the largest. IS have played a vital role in genome reduction within each species, as noted previously[95,358,387]. This chapter

demonstrates that IS load in each *Shigella* species reflects their evolutionary trajectory. *S. dysenteriae* and *S. flexneri* have similar IS loads, with no evidence of ongoing IS expansion, although IS are still actively transposing in their genomes (Figure 5.1, 5.4). This suggests that they have reached a point of IS saturation, where new transposition activity must be balanced by removal of the new insertion or an existing one by purifying selection. Within *S. dysenteriae*, all four lineages had a significant number of shared IS insertion sites, inferred to be present in the mrca, indicating that IS expansion must have occurred in the mrca of *S. dysenteriae* Sd1 (Figure 5.1). Within *S. flexneri*, each lineage had a different IS copy number (Figure 5.4, 5.5), similar to *S. sonnei*. However, in *S. flexneri*, the level of between lineage diversity is higher than either *S. sonnei* or *S. dysenteriae* (Table 5.5). The majority of *S. flexneri* lineages have been evolving for ~660 years, compared to the ~300 years of evolution across all lineages of *S. sonnei* and *S. dysenteriae* Sd1, so the expectation is that *S. flexneri* lineages will have more divergent copy numbers than *S. sonnei* and *S. dysenteriae* lineages. This is demonstrated by the high percentage of IS that are not shared between pairs of strains in different lineages in *S. flexneri* (Table 5.5).

Extant genomes in each *S. flexneri* lineage contained comparable IS copy numbers to that estimated in their mrca, indicating that like *S. dysenteriae*, most *S. flexneri* lineages have ceased IS expansion (Table 5.4). This is in contrast to *S. sonnei*, where IS copy number is still increasing, with modelling suggesting that if *S. sonnei* continues to accumulate IS at a similar trajectory, the IS saturation point would be ~380 IS copies (Chapter 4).

IS have contributed to the inactivation of hundreds of genes within *Shigella*, in addition to gene inactivation caused by missense mutations or intergenic indels. Comparison of inactivated or absent genes in *S. flexneri* and *S. sonnei* with *S. dysenteriae* revealed that both species are still undergoing genome decay (section 5.3.5.3). Both *S. flexneri* and *S. sonnei* and are on the path to more reduced genomes, however *S. dysenteriae* appears to be stabilising its genome content (section 5.3.5.3). Overall, a high number of genes were found to be inactivated in *Shigella*. It has been previously suggested that high rates of purifying selection with *Shigella* contributes to the increase in genome reduction[387]. As a small number of cells can cause a successful infection in a human host, the effective population size of *Shigella* is often much lower than the wider *E. coli* group, resulting in population bottlenecks[387].

The role of IS in the formation of pseudogenes varies across the three *Shigella* species. The rates of IS-mediated and mutational gene inactivation in *S. sonnei* is similar, and the number

of inactivated genes across *S. sonnei* and *S. dysenteriae* was similar. *S. flexneri* had many more total inactivated genes (2.5 times) than the other two species. However, individual *S. flexneri* lineages had similar numbers of inactivated genes to *S. sonnei* and *S. dysenteriae*. As previously discussed, each *S. flexneri* lineage has a similar level of diversity within them to *S. sonnei* and *S. dysenteriae* (Table 5.5), and so the high numbers of inactivated genes found in *S. flexneri* is likely due to the evolutionary distance between each lineage.

Comparison of genes that were inactive in *S. flexneri* and *S. sonnei* with gene content in *S. dysenteriae* revealed that a significant number of genes inactive in these two species were also inactive or completely absent from *S. dysenteriae* (section 5.3.5.3). These results indicate that both *S. sonnei* and *S. flexneri* are undergoing convergent gene inactivation, putting them on the same evolutionary trajectory, heading towards more reduced genomes, like *S. dysenteriae*. Given the size of *S. sonnei*'s genome, *S. sonnei* is likely still in the early stages of this genome reduction process, compared to *S. flexneri*, which already has a much smaller genome than *S. sonnei*, albeit larger than *S. dysenteriae*.

This chapter explored functional categories enriched for gene inactivation within each *Shigella* species (section 5.3.5.1, 5.3.5.2). A common theme amongst species was inactivation of carbohydrate or sugar synthesis pathways, and motility, phage or membrane genes. These pathways have been identified as prone to inactivation in other studies, but the reasons behind these inactivations are unclear[358]. Inactivation of sugar synthesis or outer membrane genes may assist with the evasion of the host immune system by *Shigella*. Another possible explanation is due to the specific niche *Shigella* inhabits, these pathways may no longer be required for survival in this environment[58].

5.4.3 IS*1* is expanded in each *Shigella* species, but not expanded in any *E. coli* lineage

Within *Shigella*, there were five IS common to all three species that contributed to the majority of IS copy number. These five IS were also found within a diverse group of *E. coli* genomes, but at much lower copy numbers than *Shigella*, and interestingly, other pathogenic lineages of *E. coli* did not show evidence of IS expansion of these five or any other IS.

The cause of IS expansion in *Shigella* is unknown, but the main driver of this expansion appears

to have been IS*1* (Figure 5.5). IS*1* is present at the highest proportion in each *Shigella* species and contributes the most to burden (Figure 5.7). There are many isoforms of IS*1*, and each isoform has small sequence differences[6]. As discussed in Chapter 1, IS*1* produces two products - InsA, and InsAB', a frameshift product of the genes *insA* and *insB*, and the ratio of these two products determines the transposition rate. InsAB' is produced by ribosomal slippage at the A_6C motif within the *insA* and *insB* genes[6]. Previous studies have shown that if this motif is modified to an A_7C or GA_2A_3C motif, then more slippage occurs, generating higher amounts of InsAB' and therefore higher transposition rates[374,375]. However, all IS*1* sequences within *Shigella* had the A_6C motif, so transposition rates have not been affected by this particular mechanism. Based on these results, it is likely that IS*1* expansion in *Shigella* may be due to purifying selection in a population that has gone through a bottleneck during host restriction or adaptation - the pathogenic *E. coli* lineages are not host restricted and so have not undergone strong purifying selection.

5.5 Summary

This chapter expands the framework established in Chapter 4 to examine IS dynamics in two additional *Shigella* species, *S. dysenteriae* and *S. flexneri*. Previous comparative genomic studies of *Shigella* revealed that each *Shigella* species has undergone genome decay, in addition to IS expansion[341]. This study has shown that the majority of this IS expansion is due to five IS, which are also common in the wider *E. coli* population, but have expanded in *Shigella* likely due to the relaxation of purifying selection. Each of three *Shigella* species are still undergoing genome decay, with ~80% of inactivated genes in *S. sonnei* and *S. flexneri* found to be inactive or completely absent in *S. dysenteriae*, the species with the most reduced genome. The results presented here indicate that each *Shigella* species is at a different point on the same evolutionary path, heading towards small, reduced genomes. Given the size of the *S. sonnei* genome, *S. sonnei* is likely at the beginning of this trajectory, and given that IS are still accumulating in *S. sonnei* genomes, IS are likely to play an important role in its genome reduction.

Chapter 6

Conclusions

6.1 Development of a novel method for examining IS in bacteria

This study presents a new tool, ISMapper, that is able to detect IS from short read data using a high-throughput approach. Prior to the development of ISMapper, there were few tools for detecting transposases from short read data, and so many large genomic studies focused on the simpler genetic variation caused by SNPs. Existing tools for transposase investigation using short read data were either aimed at detecting structural variation, including transposase mediated rearrangement, or for detecting eukaryotic transposases, frequently in humans or the fruit fly, *Drosophila*[388–392]. ISMapper was one of the first methods developed specifically for detection of transposases in bacterial populations. ISMapper was shown to be computationally fast and efficient, with a high degree of sensitivity and specificity. Since ISMapper's publication, several tools have been released to address the same problem, indicating that detection of precise insertion sites is of interest in many different fields. ISMapper was shown to be the most accurate and user-friendly of these tools, however, reference-free approaches are sometimes more suitable in some contexts (i.e. comparing IS dynamics across a diverse species such as *E. coli*).

6.1.1 ISMapper aids investigation of the maintenance and spread of AMR

To demonstrate the usefulness of ISMapper, this study applied ISMapper to explore the role of IS in the evolution of AMR within several clonal pathogen populations. IS are frequently associated with antibiotic resistance genes, and contribute to their movement and regulation. ISMapper aided the detection of antibiotic resistance regions that had transferred into new genetic contexts in high throughput genomic studies of both *S.* Typhi and *A. baumannii*. Within *S.* Typhi, the MDR transposon flanked by IS*1*, usually carried on a large plasmid, was found to have transferred into the *S.* Typhi chromosome. The IS*1* transposon had targeted four different chromosomal locations. A similar situation was found in *A. baumannii*, where a chromosomally located ISAba1 transposon was found to carry one of four different carbapenemase genes. Each lineage within the *A. baumannii* study had independently acquired the transposon, generating resistance to imipenem, and aiding *A. baumannii*'s spread throughout the hospital. Understanding the genetic context of antibiotic resistance is vital -

antibiotic resistance located in a chromosomal setting can be much more stable than the same plasmid-encoded resistance. The maintenance of extra-chromosomal plasmids generally carries a fitness cost, so plasmids can be easily lost if the selective pressure to maintain them is no longer present[276]. By maintaining the MDR transposon within the chromosome, *S.* Typhi and *A. baumannii* become much more difficult to treat. Without ISMapper, determining the location of these resistance elements would have been a more challenging task.

Movement of antibiotic resistance genes frequently introduces transposases into new genetic contexts. One particular IS, IS*26*, is commonly linked to antibiotic resistance genes[170,393]. This IS was found to be prevalent within the antibiotic resistance island, SGI, in *S.* Kentucky. Determining the different structures of the island *in silico* would not have been possible without ISMapper, as IS*26* complicated the assembly of this region. Due to its transposition mechanism, IS*26* was found to be the causative agent of the variation found within the SGI, contributing to significant rearrangements and deletions of large segments of the island. The introduction of IS*26* into the *S.* Kentucky chromosome via the SGI provides IS*26* with the ability to transpose to new locations in the chromosome, creating the possibility of additional IS*26*-mediated variation in the *S.* Kentucky genome.

Complex antibiotic resistance mechanisms that are not simply explained by the presence or absence of a gene or SNP were investigated with ISMapper. In this thesis, ISMapper was applied to a collection of highly resistant *A. baumannii* genomes, where resistance to polymyxins, the last line of defense, was emerging. Before the application of ISMapper, the causative agent of polymyxin resistance was unknown for several genomes, however, ISMapper detected IS-mediated gene inactivation of the outer membrane *lpx* genes that are known to be required for polymyxin activity. Traditional genomics approaches missed these causative mutations, and without ISMapper, more expensive long read sequencing would likely have been required to detect them.

6.1.2 Future investigative possibilities using ISMapper

The development of ISMapper has opened up new avenues of investigation for understanding the role of IS in the evolution of bacterial genomes. As has been shown in this thesis, ISMapper can be applied to a wide variety of bacterial species. Short read sequencing is still the primary method for genomic investigation in public health and hospital microbiology labs,

and the number of reads deposited into public repositories is increasing every day[394,395]. For many bacterial pathogens, there are thousands of genomes available in these repositories. ISMapper is free and open-source software that provides a high throughput method for analysing the IS in any bacterial species.

The role IS play in the evolution of many bacterial genomes is still an open question. As discussed in Chapter 1, there are several competing hypotheses surrounding the reason for IS abundance in some bacterial genomes and not others[84,85]. To date, studies have relied on a few completed genomes as representatives for whole bacterial clones or populations to draw their inferences concerning IS dynamics[53,248,341,396]. ISMapper will allow the examination of hundreds or thousands of genomes from a single population, rather than just a few representatives, improving the power of comparative genomic studies. Several such studies already exist, such as the GenomeTrackr project, which is sequencing all *Salmonella*, *Listeria* and other foodborne pathogens that enter their public health laboratories.

The advent of long read sequencing allows researchers to more easily construct completed bacterial genomes in which the detection of IS insertion sites becomes trivial, but costs are prohibitive at this scale. The ability to create new reference genomes using long read technologies will aid detection of IS from short reads using ISMapper, as ISMapper works best with a high quality close reference genome to map against. Long read sequencing also allows the confirmation of insertion sites and complex structures that IS facilitate. Combining short read sequencing data and ISMapper and long read sequencing will enable unique insights into the role IS play in bacterial evolution and antibiotic resistance.

6.2 A new framework for examining IS in *Shigella*

This thesis develops a new framework, using ISMapper, for investigating and understanding IS within three species of *Shigella*. Prior to this study, examination of IS in *Shigella* has been limited to the few completed genomes available[248,341]. Comparison of IS dynamics across species requires the examination of variation, which is not possible using the single completed genome available for *S. dysenteriae* and two genomes for *S. sonnei*. The use of ISMapper with high throughput genomic data allowed investigation of IS variation and dynamics in whole populations of bacterial species.

6.2.1 Insights into the role of IS and *Shigella* evolution

This study provided novel insights into the dynamics of IS in the evolution of *Shigella* genomes. Five IS were found to be common across all three *Shigella* species. These five IS were activity transposing in each *Shigella* species and contributed significantly to IS burden. IS*1* was the most active IS, and contributed the most to IS burden in all three *Shigella* species. In addition to identifying the dynamics of individual IS in *Shigella*, the combination of ISMapper data and previously determined population structures enabled the examination of the impact of IS on the populations of *Shigella*. *S. dysenteriae* and *S. flexneri* were found to have stable IS copy numbers, with no evidence of ongoing IS expansion, despite active IS transposition. *S. sonnei*, however, was found to have an expanding IS copy number in addition to actively transposing IS. *S. sonnei* also had a high number of inactivated genes that were either inactive or absent from *S. dysenteriae*. *S. sonnei* has the largest genome of the three *Shigella* species, indicating that IS are still playing an active role in genome reduction of this species.

Comparisons of IS in each *Shigella* species with the wider *E. coli* population revealed that *Shigella* has undergone significant IS expansion in comparison to their closest relatives. Pathogenic *E. coli* lineages were found to contain a lower IS burden than each *Shigella* species. Unlike the three *Shigella* species, IS*1* was present in the pathogenic *E. coli* lineages, but was not found to contribute to the majority of IS load. One hypothesis for this discrepancy is that control of IS*1* transposition has been released in each of the *Shigella* species, but maintained within these *E. coli* lineages. The specific control mechanism is still uncertain, as there were no obvious sequence mutations in the IS*1* sequences that could explain IS*1* expansion in each *Shigella* species.

6.2.2 Future directions for understanding the role of IS in the evolution of *Shigella*

Whilst this study answers important questions about the dynamics of IS in *Shigella*, it also opens up several additional lines of inquiry. This study investigates only three species of *Shigella*. The fourth species, *S. boydii*, is rare, and few sequenced *S. boydii* genomes exist. To date there has not been a thorough investigation of *S. boydii*'s population structure. Examination of IS burden in *S. boydii* would provide important insights into the role of IS expansion in *Shigella*, and investigate if the same five IS found in *S. sonnei*, *S. dysenteriae* and *S. flexneri* have also contributed to IS

expansion in *S. boydii*.

More broadly, this thesis raises important questions about the dynamics of the specific IS found in *Shigella*. Five IS were found across all three *Shigella* species, and these IS were also present in several *E. coli* lineages. The behaviour of these IS is different in *E. coli* compared to *Shigella*, but the mechanisms behind this remain unknown. Previous studies have examined IS in a few genomes of individual species, but ISMapper enables the investigation of the same IS in natural populations as opposed to lab-evolved populations, enabling the comparison of activity in different genetic backgrounds. An extension of the framework developed in this thesis would allow greater scrutiny of IS transposition rates in nature. Ancestral state reconstruction of IS insertion sites in *S. sonnei* examined the number of gain and loss events across the phylogeny. Previous studies have demonstrated different transposition rates for different IS in experimental settings[8,375,397]. However, to date, no studies have attempted to infer transposition rates for IS in bacterial populations evolving in nature. These three *Shigella* datasets provide a unique opportunity to investigate specific rates of gain and loss for different IS. However, inferring transposition rates from the data available would require the development of more sophisticated modelling than is attempted in this study. Development of these models would allow comparison of transposition rates for the same IS within the three different genetic backgrounds.

IS have been shown in many contexts to contribute to gene inactivation and pseudogene formation. This study only scratches the surface of the role of gene inactivation in the evolution of *Shigella*, simply identifying functional categories enriched for inactivation based on currenty available functional annotations, which are quite limited. In all three species of *Shigella*, over 40% of genes were not assigned to a functional category by RAST, so a significant amount of gene inactivation in *Shigella* is occurring in genes of unknown function. This complicates inferences about the role of gene inactivation and evolutionary adaptation. Studies are beginning to more thoroughly examine the functions of genes in *Shigella* using Tn-seq methods[398]. Some progress is being made to construct metabolic models of *E. coli*[399], however many genes are not metabolic. Additionally, these metabolic models do not yet comprehensively assess genes expressed under multiple conditions, such as interaction with hosts, other microbes, interaction with phage or different environmental conditions. Further research and tools in this area are required before a more thorough investigation of the role of IS in gene inactivation within whole populations of *Shigella*, and was outside the scope of this

thesis.

Additionally, this thesis did not examine the role of IS in gene upregulation in *Shigella*. Distance from the gene promoter to the IS insertion site are important when attempting to infer whether IS-mediated gene upregulation is occurring. Detection of an IS insertion site upstream from a gene could indicate upregulation of the downstream gene through promoter activation, or gene inactivation if the insertion site is within the gene promoter. Comparison of transcription data generated through RNA-seq or qPCR for strains with and without insertion site upstream of the gene would help identify IS insertion sites influence gene regulation. ISMapper provides the foundation for this work, but would be needed in conjunction with experimental work to confirm changes to gene expression, and is beyond the scope of this thesis.

6.3 Using ISMapper as a tool to examine IS dynamics in other bacterial pathogens

In addition to understanding the behaviour of specific IS and their role in genome decay in *Shigella*, this study presents a novel framework for the investigation of IS dynamics in other bacterial pathogens. Here I compared *Shigella* species to other *E. coli*, however there are several other bacterial pathogens where IS have played a similar role in genome evolution as they have in *Shigella* species, which would make interesting comparisons. These include *B. mallei*[53], *Y. pestis*[60], *M. leprae*[400], *Leptospira borgpetersenii*[401], *Mycobacterium ulcerans*[402] and *B. pertussis*[92], where IS have contributed to genome reduction and host adaptation. The brief analysis of IS*6110* in 138 *M. tuberculosis* genomes in Chapter 2 highlights that this framework can be applied to other pathogens to examine lineage differences and insertion hotspots.

IS have also facilitated the adaptation of bacterial pathogens to new environmental niches. IS*16* has contributed to genome plasticity in *E. faecium*, promoting its adaptation to a hospital niche[403]. Studies of IS diversity in *A. baumannii* show that the most common insertion sites for ISAba1 in *A. baumannii* genomes are upstream of *ampC*, influencing carbapenem resistance, and upstream of $bla_{OXA\text{-}51}$, resulting in resistance to meropenem and imipenem[269]. Within *K. pneumoniae*, the most common IS insertions are ISKpn6 and ISKpn7, which are frequently found in the transposon Tn*4401* that carries the bla_{KPC} gene, contributing to the spread of antibiotic resistance in this pathogen[269]. Both of these species are hospital acquired

pathogens, and these mutations conferring additional antibiotic resistance contributes to their burden in hospitals. Overall, these examples only begin to comprehend the vast contribution of IS to bacterial genome evolution, and greater research is required to uncover the different IS dynamics in different bacterial species.

References

1. Achtman, M. Evolution, population structure, and phylogeography of genetically monomorphic bacterial pathogens. *Annual Review of Microbiology* **62,** 53–70 (2008).

2. Griffiths, A. J., Miller, J. H., Suzuki, D. T., Lewontin, R. C. & Gelbart, W. M. in *An introduction to genetic analysis* (W. H. Freeman, 2000).

3. Majewski, J. & Cohan, F. M. Adapt globally, act locally: the effect of selective sweeps on bacterial sequence diversity. *Genetics* **152,** 1459–1474 (1999).

4. Normark, B. H. & Normark, S. Evolution and spread of antibiotic resistance. *Journal of Internal Medicine* **252,** 91–106 (2002).

5. Kleckner, N. Transposable elements in prokaryotes. *Annual Review of Genetics* **15,** 341–404 (1981).

6. Mahillon, J. & Chandler, M. Insertion Sequences. *Microbiology and Molecular Biology Reviews* **62,** 725–774 (1998).

7. Lupski, J. R. Molecular mechanisms for transposition of drug-resistance genes and other movable genetic elements. *Review of Infectious Diseases* **9,** 357–368 (1987).

8. Kleckner, N. Regulation of transposition in bacteria. *Annual Review of Cell Biology* **6,** 297–327 (1990).

9. Nagy, Z. & Chandler, M. Regulation of transposition in bacteria. *Research in Microbiologoy* **155,** 387–398 (2004).

10. Craig, N. L. Tn*7*: a target site-specific transposon. *Molecular Microbiology* **5,** 2569–2573 (1991).

11. Betteridge, T., Partridge, S. R., Iredell, J. R. & Stokes, H. W. Genetic context and structural diversity of class 1 integrons from human commensal bacteria in a hospital intensive care unit. *Antimicrobial Agents and Chemotherapy* **55,** 3939–3943 (2011).

12. Bennett, P. M. Integrons and gene cassettes: a genetic construction kit for bacteria. *Journal*

of Antimicrobial Chemotherapy (1999).

13. Fluit, A. C. & Schmitz, F. J. Resistance integrons and super-integrons. *Clinical Microbiology and Infection* **10,** 272–288 (2004).

14. Leverstein-van Hall, M. A. *et al.* Evidence of extensive interspecies transfer of integron-mediated antimicrobial resistance genes among multidrug-resistant Enterobacteriaceae in a clinical setting. *Journal of Infectious Diseases* **186,** 49–56 (2002).

15. Marshall, B. M., Ochieng, D. J. & Levy, S. B. Commensals: underappreciated reservoir of antibiotic resistance. *Microbe* **4,** (2009).

16. Hall, R. M. & Collis, C. M. Mobile gene cassettes and integrons: capture and spread of genes by site-specific recombination. *Molecular Microbiology* **15,** 593–600 (1995).

17. Stokes, H. W. & Hall, R. M. A novel family of potentially mobile DNA elements encoding site-specific gene-integration functions: integrons. *Molecular Microbiology* (1989).

18. Hall, R. M. Integrons and gene cassettes: hotspots of diversity in bacterial genomes. *Annals of the New York Academy of Sciences* **1267,** 71–78 (2012).

19. Recchia, G. D. & Hall, R. M. Gene cassettes: a new class of mobile element. *Microbiology* (1995).

20. Frost, L. S., Leplae, R., Summers, A. O. & Toussaint, A. Mobile genetic elements: the agents of open source evolution. *Nature Reviews Microbiology* **3,** 722–732 (2005).

21. Lwoff, A. Lysogeny. *Bacteriological Reviews* **17,** 269–337 (1953).

22. Zinder, N. D. & Lederberg, J. Genetic exchange in *Salmonella*. *Journal of Bacteriology* **64,** 679–699 (1952).

23. Hacker, J. & Kaper, J. B. Pathogenicity islands and the evolution of microbes. *Annual Review of Microbiology* **54,** 641–679 (2000).

24. Juhas, M. *et al.* Genomic islands: tools of bacterial horizontal gene transfer and evolution. *FEMS Microbiology Reviews* **33,** 376–393 (2009).

25. Ravatn, R., Studer, S., Springael, D., Zehnder, A. J. & Meer, J. R. van der. Chromosomal integration, tandem amplification, and deamplification in *Pseudomonas putida* F1 of a

105-kilobase genetic element containing the chlorocatechol degradative genes from *Pseudomonas sp.* strain B13. *Journal of Bacteriology* **180,** 4360–4369 (1998).

26. Meer, J. R. van der, Ravatn, R. & Sentchilo, V. The clc element of *Pseudomonas sp.* strain B13 and other mobile degradative elements employing phage-like integrases. *Archives of Microbiology* **175,** 79–85 (2001).

27. Mohd-Zain, Z. *et al.* Transferable antibiotic resistance elements in *Haemophilus influenzae* share a common evolutionary origin with a diverse family of syntenic genomic islands. *Journal of Bacteriology* **186,** 8114–8122 (2004).

28. Schmidt, H. & Hensel, M. Pathogenicity islands in bacterial pathogenesis. *Clinical Microbiology Reviews* **17,** 14–56 (2004).

29. Burrus, V. & Waldor, M. K. Control of SXT integration and excision. *Journal of Bacteriology* **185,** 5045–5054 (2003).

30. Shoemaker, N. B., Wang, G. R. & Salyers, A. A. Multiple gene products and sequences required for excision of the mobilizable integrated *Bacteroides* element NBU1. *Journal of Bacteriology* **182,** 928–936 (2000).

31. Hochhut, B., Marrero, J. & Waldor, M. K. Mobilization of plasmids and chromosomal DNA mediated by the SXT element, a constin found in *Vibrio cholerae* O139. *Journal of Bacteriology* **182,** 2043–2047 (2000).

32. Hacker, J., Blum-Oehler, G., Mühldorfer, I. & Tschäpe, H. Pathogenicity islands of virulent bacteria: structure, function and impact on microbial evolution. *Molecular Microbiology* **23,** 1089–1097 (1997).

33. Dobrindt, U., Hochhut, B., Hentschel, U. & Hacker, J. Genomic islands in pathogenic and environmental microorganisms. *Nature Reviews Microbiology* **2,** 414–424 (2004).

34. Vernikos, G. S., Thomson, N. R. & Parkhill, J. Genetic flux over time in the *Salmonella* lineage. *Genome Biology* **8,** R100 (2007).

35. Gehring, A. M. *et al.* Iron acquisition in plague: modular logic in enzymatic biogenesis of yersiniabactin by *Yersinia pestis*. *Chemistry & Biology* **5,** 573–586 (1998).

36. Bach, S., Almeida, A. de & Carniel, E. The *Yersinia* high-pathogenicity island is present in different members of the family Enterobacteriaceae. *FEMS Microbiology Letters* **183,** 289–294

(2000).

37. Baar, C. *et al.* Complete genome sequence and analysis of *Wolinella succinogenes*. *Proceedings of the National Academy of Sciences of the United States of America* **100,** 11690–11695 (2003).

38. Boyd, D. *et al.* Complete nucleotide sequence of a 43-kilobase genomic island associated with the multidrug resistance region of *Salmonella enterica* serovar Typhimurium DT104 and its identification in phage type DT120 and serovar Agona. *Journal of Bacteriology* **183,** 5725–5732 (2001).

39. Ahmed, A. M., Hussein, A. I. A. & Shimamoto, T. *Proteus mirabilis* clinical isolate harbouring a new variant of *Salmonella* genomic island 1 containing the multiple antibiotic resistance region. *Journal of Antimicrobial Chemotherapy* **59,** 184–190 (2007).

40. Luck, S. N., Turner, S. A., Rajakumar, K., Adler, B. & Sakellaris, H. Excision of the *Shigella* resistance locus pathogenicity island in *Shigella flexneri* is stimulated by a member of a new subgroup of recombination directionality factors. *Journal of Bacteriology* **186,** 5551–5554 (2004).

41. Beaber, J. W., Hochhut, B. & Waldor, M. K. Genomic and functional analyses of SXT, an integrating antibiotic resistance gene transfer element derived from *Vibrio cholerae*. *Journal of Bacteriology* **184,** 4259–4269 (2002).

42. Zhang, R. & Zhang, C.-T. Identification of genomic islands in the genome of *Bacillus cereus* by comparative analysis with *Bacillus anthracis*. *Physiological Genomics* **16,** 19–23 (2003).

43. Baker, S. *et al.* A novel linear plasmid mediates flagellar variation in *Salmonella* Typhi. *PLoS Pathogens* **3,** e59 (2007).

44. Couturier, M., Bex, F., Bergquist, P. L. & Maas, W. K. Identification and classification of bacterial plasmids. *Microbiological Reviews* **52,** 375–395 (1988).

45. Datta, N. & Hedges, R. W. Compatibility groups among fi - R factors. *Nature* **234,** 222–223

(1971).

46. Frost, L. S. *Conjugation*. (Plenum, 1993).

47. Wilkins, B. M. & Frost, L. S. *Molecular Medical Microbiology*. (Acadmeic, 2001).

48. Helsinki, D. R. *The Horizontal Gene Pool: Bacterial Plasmid and Gene Spread*. (Harwood Academic, 2000).

49. Gómez-Lus, R. Evolution of bacterial resistance to antibiotics during the last three decades. *International Microbiology* **1,** 279–284 (1998).

50. Welch, T. J. *et al.* Multiple antimicrobial resistance in plague: an emerging public health risk. *PLoS ONE* **2,** e309 (2007).

51. Parkhill, J. *et al.* Complete genome sequence of a multiple drug resistant *Salmonella enterica* serovar Typhi CT18. *Nature* **413,** 848–852 (2001).

52. Holt, K. E. *et al.* Emergence of a globally dominant IncHI1 plasmid type associated with multiple drug resistant typhoid. *PLoS Neglected Tropical Diseases* **5,** e1245 (2011).

53. Losada, L. *et al.* Continuing evolution of *Burkholderia mallei* through genome reduction and large-scale rearrangements. *Genome Biology and Evolution* **2,** 102–116 (2010).

54. Moran, N. A. Microbial minimalism: genome reduction in bacterial pathogens. *Cell* **108,** 583–586 (2002).

55. Andersson, J. O. & Andersson, S. G. Pseudogenes, junk DNA, and the dynamics of *Rickettsia* genomes. *Molecular Biology and Evolution* **18,** 829–839 (2001).

56. Mira, A., Ochman, H. & Moran, N. A. Deletional bias and the evolution of bacterial genomes. *Trends in Genetics* (2001).

57. Andersson, S. G. E. & Kurland, C. G. Reductive evolution of resident genomes. *Trends in Microbiology* **6,** 263–268 (1998).

58. Moran, N. A. & Wernegreen, J. J. Lifestyle evolution in symbiotic bacteria: insights from genomics. *Trends in Ecology & Evolution* **15,** 321–326 (2000).

59. Parkhill, J. & Thomson, N. Evolutionary strategies of human pathogens. *Cold Spring Harbor*

Symposia on Quantitative Biology **68,** 151–158 (2003).

60. Parkhill, J. *et al.* Genome sequence of *Yersinia pestis*, the causative agent of plague. *Nature* **413,** 523–527 (2001).

61. Thomson, N. R. *et al.* Comparative genome analysis of *Salmonella* Enteritidis PT4 and *Salmonella* Gallinarum 287/91 provides insights into evolutionary and host adaptation pathways. *Genome Research* **18,** 1624–1637 (2008).

62. Nuccio, S.-P. & Baumler, A. J. Comparative analysis of *Salmonella* genomes identifies a metabolic network for escalating growth in the inflamed gut. *mBio* **5,** e00929–14 (2014).

63. Holt, K. E. *et al.* Pseudogene accumulation in the evolutionary histories of *Salmonella* enterica serovars Paratyphi A and Typhi. *BMC Genomics* **10,** 36 (2009).

64. Vos, M. Why do bacteria engage in homologous recombination? *Trends in Microbiology* **17,** 226–232 (2009).

65. Bennett, P. M. Genome plasticity: insertion sequence elements, transposons and integrons, and DNA rearrangement. *Methods in Molecular Biology* **266,** 71–113 (2004).

66. Thomas, C. M. & Nielsen, K. M. Mechanisms of, and barriers to, horizontal gene transfer between bacteria. *Nature Reviews Microbiology* **3,** 711–721 (2005).

67. Robinson, D. A. & Enright, M. C. Evolution of *Staphylococcus aureus* by large chromosomal replacements. *Journal of Bacteriology* **186,** 1060–1064 (2004).

68. Wyres, K. L. *et al.* Extensive capsule locus variation and large-scale genomic recombination within the *Klebsiella pneumoniae* clonal group 258. *Genome Biology and Evolution* **7,** 1267–1279 (2015).

69. Schultz, M. B. *et al.* Repeated local emergence of carbapenem resistant *Acinetobacter baumannii* in a single hospital ward. *Microbial Genomics* (2016).

70. Croucher, N. J. *et al.* Population genomics of post-vaccine changes in pneumococcal epidemiology. *Nature Genetics* **45,** 656–663 (2013).

71. Bryant, J., Chewapreecha, C. & Bentley, S. D. Developing insights into the mechanisms of evolution of bacterial pathogens from whole-genome sequences. *Future Microbiology* **7,** 1283–

1296 (2012).

72. Siguier, P., Gourbeyre, E., Varani, A., Ton-Hoang, B. & Chandler, M. Everyman's guide to bacterial insertion sequences. *Microbiology Spectrum* **3,** (2015).

73. Lawrence, J. G., Ochman, H. & Hartl, D. L. The evolution of insertion sequences within enteric bacteria. *Genetics* **131,** 9–20 (1992).

74. Biserčić, M. & Ochman, H. The ancestry of insertion sequences common to *Escherichia coli* and *Salmonella typhimurium*. *Journal of Bacteriology* **175,** 7863–7868 (1993).

75. Siguier, P., Filée, J. & Chandler, M. Insertion sequences in prokaryotic genomes. *Current Opinion in Microbiology* **9,** 526–531 (2006).

76. Sawyer, S. A. *et al.* Distribution and abundance of insertion sequences among natural isolates of *Escherichia coli*. *Genetics* **115,** 51–63 (1987).

77. Soria, G., Barbé, J. & Gibert, I. Molecular fingerprinting of *Salmonella typhimurium* by IS*200*-typing as a tool for epidemiological and evolutionary studies. *Microbiología* **10,** 57–68 (1994).

78. Soldati, L. & Piffaretti, J. C. Molecular typing of *Shigella* strains using pulsed field gel electrophoresis and genome hybridization with insertion sequences. *Research in Microbiologoy* **142,** 489–498 (1991).

79. Cave, M. D., Eisenach, K. D., McDermott, P. F., Bates, J. H. & Crawford, J. T. IS*6110*: Conservation of sequence in the *Mycobacterium tuberculosis* complex and its utilization in DNA fingerprinting. *Molecular and Cellular Probes* **5,** 73–80 (1991).

80. Das, S., Paramasivan, C. N., Lowrie, D. B., Prabhakar, R. & Narayanan, P. R. IS*6110* restriction fragment length polymorphism typing of clinical isolates of *Mycobacterium tuberculosis* from patients with pulmonary tuberculosis in Madras, South India. *Tubercle and Lung Disease* **76,** 550–554 (1995).

81. Voskresenskaya, E., Savin, C., Leclercq, A., Tseneva, G. & Carniel, E. Typing and clustering of *Yersinia pseudotuberculosis* isolates by restriction fragment length polymorphism analysis using insertion sequenes. *Journal of Clinical Microbiology* **52,** 1978–1989 (2014).

82. Werner, G. *et al.* IS element IS*16* as a molecular screening tool to identify hospital-associated

strains of *Enterococcus faecium*. *BMC Infectious Diseases* **11,** 80 (2011).

83. Tanaka, M. M., Rosenberg, N. A. & Small, P. M. The control of copy number of IS*6110* in *Mycobacterium tuberculosis*. *Molecular Biology and Evolution* **21,** 2195–2201 (2004).

84. Touchon, M. & Rocha, E. P. C. Causes of insertion sequences abundance in prokaryotic genomes. *Molecular Biology and Evolution* **24,** 969–981 (2007).

85. Iranzo, J., Gómez, M. J., Saro, F. J. L. de & Manrubia, S. Large-scale genomic analysis suggests a neutral punctuated dynamics of transposable elements in bacterial genomes. *PLoS Computational Biology* **10,** e1003680 (2014).

86. Wu, Y., Aandahl, R. Z. & Tanaka, M. M. Dynamics of bacterial insertion sequences: can transposition bursts help the elements persist? *BMC Evolutionary Biology* **15,** 1 (2015).

87. Naas, T., Blot, M., Fitch, W. M. & Arber, W. Dynamics of IS-related genetic rearrangements in resting *Escherichia coli* K-12. *Molecular Biology and Evolution* **12,** 198–207 (1995).

88. Papadopoulos, D. *et al.* Genomic evolution during a 10,000-generation experiment with bacteria. *Proceedings of the National Academy of Sciences of the United States of America* **96,** 3807–3812 (1999).

89. Schneider, D. & Lenski, R. E. Dynamics of insertion sequence elements during experimental evolution of bacteria. *Research in Microbiologoy* **155,** 319–327 (2004).

90. Sousa, A., Bourgard, C., Wahl, L. M. & Gordo, I. Rates of transposition in *Escherichia coli*. *Biology Letters* **9,** 20130838–20130838 (2013).

91. Singh, P. & Cole, S. T. *Mycobacterium leprae*: genes, pseudogenes and genetic diversity. *Future Microbiology* **6,** 57–71 (2011).

92. Parkhill, J. *et al.* Comparative analysis of the genome sequences of *Bordetella pertussis*, *Bordetella parapertussis* and *Bordetella bronchiseptica*. *Nature Genetics* **35,** 32–40 (2003).

93. Song, H. *et al.* The early stage of bacterial genome-reductive evolution in the host. *PLoS Pathogens* **6,** e1000922 (2010).

94. Ooka, T. *et al.* Inference of the impact of insertion sequence (IS) elements on bacterial genome diversification through analysis of small-size structural polymorphisms in *Escherichia*

coli O157 genomes. *Genome Research* **19,** 1809–1816 (2009).

95. Prosseda, G. *et al.* Shedding of genes that interfere with the pathogenic lifestyle: the *Shigella* model. *Research in Microbiology* **163,** 399–406 (2012).

96. So, M., Heffron, F. & McCarthy, B. J. The *E. coli* gene encoding heat stable toxin is a bacterial transposon flanked by inverted repeats of IS*1*. *Nature* **277,** 453–456 (1979).

97. So, M. & McCarthy, B. J. Nucleotide sequence of the bacterial transposon Tn*1681* encoding a heat-stable (ST) toxin and its identification in enterotoxigenic *Escherichia coli* strains. *Proceedings of the National Academy of Sciences of the United States of America* **77,** 4011–4015 (1980).

98. Alton, N. K. & Vapnek, D. Nucleotide sequence analysis of the chloramphenicol resistance transposon Tn*9*. *Nature* **282,** 864–869 (1979).

99. Campos, J. C. *et al. Antimicrobial Agents and Chemotherapy* **59,** 7387–7395 (2015).

100. Hamidian, M., Holt, K. E., Pickard, D., Dougan, G. & Hall, R. M. A GC1 *Acinetobacter baumannii* isolate carrying AbaR3 and the aminoglycoside resistance transposon Tn*aphA6* in a conjugative plasmid. *Journal of Antimicrobial Chemotherapy* **69,** 955–958 (2014).

101. Holt, K. E. *et al.* Multidrug-resistant *Salmonella enterica* serovar Paratyphi A harbors IncHI1 plasmids similar to those found in serovar Typhi. *Journal of Bacteriology* **189,** 4257–4264 (2007).

102. Harmer, C. J. & Hall, R. M. *Microbial Drug Resistance* **20,** 416–423 (2014).

103. Sheppard, A. E. *et al.* Nested Russian Doll-like genetic mobility drives rapid dissemination of the carbapenem resistance gene bla_{KPC}. *Antimicrobial Agents and Chemotherapy* **60,** 3767–3778 (2016).

104. Post, V., White, P. A. & Hall, R. M. Evolution of AbaR-type genomic resistance islands in multiply antibiotic-resistant *Acinetobacter baumannii*. *Journal of Antimicrobial Chemotherapy* **65,** 1162–1170 (2010).

105. Safi, H. *et al.* IS*6110* functions as a mobile, monocyte-activated promoter in *Mycobacterium tuberculosis*. *Molecular Microbiology* **52,** 999–1012 (2004).

106. Soto, C. Y. *et al.* IS*6110* mediates increased transcription of the *phoP* virulence gene in

a multidrug-resistant clinical isolate responsible for tuberculosis outbreaks. *Journal of Clinical Microbiology* **42,** 212–219 (2004).

107. Uria, M. J. *et al.* A generic mechanism in *Neisseria meningitidis* for enhanced resistance against bactericidal antibodies. *Journal of Experimental Medicine* **205,** 1423–1434 (2008).

108. Coleman, N. V., Richardson-Harris, J., Wilson, N. L. & Holmes, A. J. Insertion sequence ISPst4 activates pUC plasmid replication in *Pseudomonas stutzeri*. *FEMS Microbiology Letters* **356,** 242–249 (2014).

109. Aronson, B. D., Levinthal, M. & Somerville, R. L. Activation of a cryptic pathway for threonine metabolism via specific IS*3*-mediated alteration of promoter structure in *Escherichia coli*. *Journal of Bacteriology* **171,** 5503–5511 (1989).

110. Hall, B. G. Activation of the *bgl* operon by adaptive mutation. *Molecular Biology and Evolution* **15,** 1–5 (1998).

111. Van Der Ploeg, J., Willemsen, M., Van Hall, G. & Janssen, D. B. Adaptation of *Xanthobacter autotrophicus* GJ10 to bromoacetate due to activation and mobilization of the haloacetate dehalogenase gene by insertion element IS*1247*. *Journal of Bacteriology* **177,** 1348–1356 (1995).

112. Džidić, S., Šušković, J. & Kos, B. Antibiotic resistance mechanisms in bacteria: biochemical and genetic aspects. *Food Technology and Biotechnology* **46,** 11–21 (2008).

113. Giedraitienė, A., Vitkauskienė, A., Naginienė, R. & Pavilonis, A. Antibiotic resistance mechanisms of clinically important bacteria. *Medicina* **47,** 137–146 (2011).

114. Li, X.-Z. & Nikaido, H. Efflux-mediated drug resistance in bacteria: an update. *Drugs* **69,** 1555–1623 (2009).

115. Foley, S. L. & Lynne, A. M. Food animal-associated *Salmonella* challenges: pathogenicity and antimicrobial resistance. *Journal of Animal Science* **86,** E173–87 (2008).

116. Alekshun, M. N. & Levy, S. B. Molecular mechanisms of antibacterial multidrug resistance. *Cell* **128,** 1037–1050 (2007).

117. Okamoto, K., Gotoh, N. & Nishino, T. *Pseudomonas aeruginosa* reveals high intrinsic resistance to penem antibiotics: penem resistance mechanisms and their interplay.

Antimicrobial Agents and Chemotherapy **45,** 1964–1971 (2001).

118. Olaitan, A. O., Morand, S. & Rolain, J.-M. Mechanisms of polymyxin resistance: acquired and intrinsic resistance in bacteria. *Frontiers in Microbiology* **5,** 643 (2014).

119. Piddock, L. J. V. Multidrug-resistance efflux pumps - not just for resistance. *Nature Reviews Microbiology* **4,** 629–636 (2006).

120. Perron, G. G. *et al.* Functional characterization of bacteria isolated from ancient arctic soil exposes diverse resistance mechanisms to modern antibiotics. *PLoS ONE* **10,** e0069533 (2015).

121. Miriagou, V., Carattoli, A. & Fanning, S. Antimicrobial resistance islands: resistance gene clusters in *Salmonella* chromosome and plasmids. *Microbes and Infection* **8,** 1923–1930 (2006).

122. Bailey, J. K., Pinyon, J. L., Anantham, S. & Hall, R. M. Distribution of the bla_{TEM} gene and bla_{TEM}-containing transposons in commensal *Escherichia coli*. *Journal of Antimicrobial Chemotherapy* **66,** 745–751 (2011).

123. Carattoli, A. Resistance plasmid families in Enterobacteriaceae. *Antimicrobial Agents and Chemotherapy* **53,** 2227–2238 (2009).

124. Bennett, P. M. Plasmid encoded antibiotic resistance: acquisition and transfer of antibiotic resistance genes in bacteria. *British Journal of Pharmacology* **153,** S347–57 (2008).

125. Olaitan, A. O. *et al.* Worldwide emergence of colistin resistance in *Klebsiella pneumoniae* from healthy humans and patients in Lao PDR, Thailand, Israel, Nigeria and France owing to inactivation of the PhoP/PhoQ regulator *mgrB*: an epidemiological and molecular study. *International Journal of Antimicrobial Agents* 1–8 (2014).

126. Machida, C., Machida, Y. & Ohtsubo, E. Both inverted repeat sequences located at the ends of IS*1* provide promoter functions. *Journal of Molecular Biology* **177,** 247–267 (1984).

127. Kamruzzaman, M. *et al.* Relative strengths of promoters provided by common mobile genetic elements associated with resistance gene expression in Gram-negative bacteria. *Antimicrobial Agents and Chemotherapy* **59,** 5088–5091 (2015).

128. Jellen-Ritter, A. S. & Kern, W. V. Enhanced expression of the multidrug efflux pumps AcrAB and AcrEF associated with insertion element transposition in *Escherichia coli* mutants selected

with a fluoroquinolone. *Antimicrobial Agents and Chemotherapy* **45,** 1467–1472 (2001).

129. Lopes, B. S. & Amyes, S. G. B. Insertion sequence disruption of *adeR* and ciprofloxacin resistance caused by efflux pumps and *gyrA* and *parC* mutations in *Acinetobacter baumannii*. *International Journal of Antimicrobial Agents* **41,** 117–121 (2013).

130. Hamidian, M. & Hall, R. M. ISAba1 targets a specific position upstream of the intrinsic *ampC* gene of *Acinetobacter baumannii* leading to cephalosporin resistance. *Journal of Antimicrobial Chemotherapy* **68,** 2682–2683 (2013).

131. Hamidian, M., Hancock, D. P. & Hall, R. M. Horizontal transfer of an ISAba125-activated *ampC* gene between *Acinetobacter baumannii* strains leading to cephalosporin resistance. *Journal of Antimicrobial Chemotherapy* **68,** 244–245 (2013).

132. Zander, E., Seifert, H. & Higgins, P. G. Insertion sequence IS*18* mediates overexpression of $bla_{OXA-257}$ in a carbapenem-resistant *Acinetobacter bereziniae* isolate. *Journal of Antimicrobial Chemotherapy* **69,** 270–271 (2014).

133. Mizuuchi, K. Transpositional recombination: mechanistic insights from studies of Mu and other elements. *Annual Review of Biochemistry* **61,** 1011–1051 (1992).

134. Hickman, A. B. *et al.* DNA recognition and the precleavage state during single-stranded DNA transposition in *D. radiodurans*. *The EMBO Journal* **29,** 3840–3852 (2010).

135. Choi, S., Ohta, S. & Ohtsubo, E. A novel IS element, IS*621*, of the IS*110*/IS*492* family transposes to a specific site in repetitive extragenic palindromic sequences in *Escherichia coli*. *Journal of Bacteriology* **185,** 4891–4900 (2003).

136. Siguier, P., Gourbeyre, E. & Chandler, M. Bacterial insertion sequences: their genomic impact and diversity. *FEMS Microbiology Reviews* **38,** 865–891 (2014).

137. Chandler, M. *et al.* Breaking and joining single-stranded DNA: the HUH endonuclease superfamily. *Nature Reviews Microbiology* **11,** 525–538 (2013).

138. Del Pilar Garcillán Barcia, M., Bernales, I., Mendiola, M. V. & De La Cruz, F. Single-stranded DNA intermediates in IS*91* rolling-circle transposition. *Molecular Microbiology* **39,** 494–502 (2001).

139. Hoang, B. T. *et al.* Transposition of ISHp608, member of an unusual family of bacterial

insertion sequences. *The EMBO Journal* **24,** 3325–3338 (2005).

140. Mendiola, M. V. & De La Cruz, F. IS*91* transposase is related to the rolling-circle-type replication proteins of the pUB110 family of plasmids. *Nucleic Acids Research* **20,** 3521–3521 (1992).

141. Hickman, A. B. & Dyda, F. Mechanisms of DNA Transposition. *Microbiology Spectrum* **3,** (2015).

142. Grindley, N. D. F. in *Mobile dna ii* (eds. Craig, N. L., Craigie, R., Gellert, M. & Lambowitz, A. M.) 272–304 (American Society of Microbiology, 2002).

143. Filée, J., Siguier, P. & Chandler, M. I am what I eat and I eat what I am: acquisition of bacterial genes by giant viruses. *Trends in Genetics* (2007).

144. Chandler, M. & Fayet, O. Translational frameshifting in the control of transposition in bacteria. *Molecular Microbiology* **7,** 497–503 (1993).

145. Ma, C. & Simons, R. W. The IS*10* antisense RNA blocks ribosome binding at the transposase translation initiation site. *The EMBO Journal* **9,** 1267–1274 (1990).

146. Kretschmer, P. J. & Cohen, S. N. Effect of temperature on translocation frequency of the Tn*3* element. *Journal of Bacteriology* **139,** 515–519 (1979).

147. Haren, L., Bétermier, M., Polard, P. & Chandler, M. IS*911*-mediated intramolecular transposition is naturally temperature sensitive. *Molecular Microbiology* **25,** 531–540 (1997).

148. Shiga, Y., Sekine, Y., Kano, Y. & Ohtsubo, E. Involvement of H-NS in transpositional recombination mediated by IS*1*. *Journal of Bacteriology* **183,** 2476–2484 (2001).

149. Gamas, P., Chandler, M. G., Prentki, P. & Galas, D. J. *Escherichia coli* integration host factor binds specifically to the ends of the insertion sequence IS*1* and to its major insertion hot-spot in pBR322. *Journal of Molecular Biology* **195,** 261–272 (1987).

150. Sewitz, S., Crellin, P. & Chalmers, R. The positive and negative regulation of Tn*10* transposition by IHF is mediated by structurally asymmetric transposon arms. *Nucleic Acids Research* **31,** 5868–5876 (2003).

151. C Turlan, M. C. IS*1*-mediated intramolecular rearrangements: formation of excised

transposon circles and replicative deletions. *The EMBO Journal* **14,** 5410–57 (1995).

152. Duval Valentin, G. & Chandler, M. Cotranslational control of DNA transposition: a window of opportunity. *Molecular Cell* **44,** 989–996 (2011).

153. Hickman, A. B., Chandler, M. & Dyda, F. Integrating prokaryotes and eukaryotes: DNA transposases in light of structure. *Critical Reviews in Biochemistry and Molecular Biology* **45,** 50–69 (2010).

154. De Palmenaer, D., Siguier, P. & Mahillon, J. IS*4* family goes genomic. *BMC Evolutionary Biology* **8,** 18 (2008).

155. Klenchin, V. A. *et al.* Phosphate coordination and movement of DNA in the Tn*5* synaptic complex: role of the (R)YREK motif. *Nucleic Acids Research* **36,** 5855–5862 (2008).

156. Kleckner, N., Chalmers, R. M., Kwon, D., Sakai, J. & Bolland, S. in *Transposable elements* 49–82 (Springer Berlin Heidelberg, 1996).

157. Reznikoff, W. S. The Tn*5* transposon. *Annual Reviews in Microbiology* (1993).

158. Reimmann, C. & Haas, D. The *istA* gene of insertion sequence IS*21* is essential for cleavage at the inner 3' ends of tandemly repeated IS*21* elements in vitro. *The EMBO Journal* **9,** 4055–4063 (1990).

159. Han, C. G., Shiga, Y., Tobe, T., Sasakawa, C. & Ohtsubo, E. Structural and functional characterization of IS*679* and IS*66*-family elements. *Journal of Bacteriology* **183,** 4296–4304 (2001).

160. Tenzen, T. & Ohtsubo, E. Preferential transposition of an IS*630*-associated composite transposon to TA in the 5'-CTAG-3' sequence. *Journal of Bacteriology* **173,** 6207–6212 (1991).

161. Williams, K., Doak, T. G. & Herrick, G. Developmental precise excision of *Oxytricha trifallax* telomere-bearing elements and formation of circles closed by a copy of the flanking target duplication. *The EMBO Journal* **12,** 4593–4601 (1993).

162. Higgins, B. P., Carpenter, C. D. & Karls, A. C. Chromosomal context directs high-frequency precise excision of IS*492* in *Pseudoalteromonas atlantica*. *Proceedings of the National Academy of Sciences of the United States of America* **104,** 1901–1906 (2007).

163. Perkins-Balding, D., Duval-Valentin, G. & Glasgow, A. C. Excision of IS*492* requires

flanking target sequences and results in circle formation in *Pseudoalteromonas atlantica*. *Journal of Bacteriology* **181,** 4937–4948 (1999).

164. Tetu, S. G. & Holmes, A. J. A family of insertion sequences that impacts integrons by specific targeting of gene cassette recombination sites, the IS*1111-attC* Group. *Journal of Bacteriology* **190,** 4959–4970 (2008).

165. Lauf, U., Müller, C. & Herrmann, H. Identification and characterisation of IS*1383*, a new insertion sequence isolated from *Pseudomonas putida* strain H. *FEMS Microbiology Letters* **170,** 407–412 (1999).

166. Müller, C., Lauf, U. & Hermann, H. The inverted repeats of IS*1384*, a newly described insertion sequence from *Pseudomonas putida* strain H, represent the specific target for integration of IS*1383*. *Molecular Genetics and Genomics* **265,** 1004–1010 (2001).

167. Kersulyte, D., Akopyants, N. S., Clifton, S. W., Roe, B. A. & Berg, D. E. Novel sequence organization and insertion specificity of IS*605* and IS*606*: chimaeric transposable elements of *Helicobacter pylori*. *Gene* **223,** 175–186 (1998).

168. He, S. *et al.* The IS*200*/IS*605* family and 'peel and paste' single-strand transposition mechanism. *Microbiology Spectrum* **3,** (2015).

169. Simpson, A. E., Skurray, R. A. & Firth, N. An IS*257*-derived hybrid promoter directs rranscription of a *tetA*(K) tetracycline resistance gene in the *Staphylococcus aureus* chromosomal *mec* region. *Journal of Bacteriology* **182,** 3345–3352 (2000).

170. Doublet, B., Praud, K., Weill, F.-X. & Cloeckaert, A. Association of IS*26*-composite transposons and complex In4-type integrons generates novel multidrug resistance loci in *Salmonella* genomic island 1. *Journal of Antimicrobial Chemotherapy* **63,** 282–289 (2009).

171. Nigro, S. J., Farrugia, D. N., Paulsen, I. T. & Hall, R. M. A novel family of genomic resistance islands, AbGRI2, contributing to aminoglycoside resistance in Acinetobacter baumannii isolates belonging to global clone 2. *Journal of Antimicrobial Chemotherapy* **68,** 554–557 (2013).

172. Bertini, A. *et al. Antimicrobial Agents and Chemotherapy* **51,** 2324–2328 (2007).

173. Loman, N. J. *et al.* Performance comparison of benchtop high-throughput sequencing

platforms. *Nature Biotechnology* **30,** 434–439 (2012).

174. Wick, Ryan R, Judd, Louise M, Gorrie, Claire L & Holt, K. E. Completing bacterial genome assemblies with multiplex MinION sequencing. *bioRxiv* (2017).

175. Pevzner, P. A., Tang, H. & Waterman, M. S. An Eulerian path approach to DNA fragment assembly. *Proceedings of the National Academy of Sciences of the United States of America* **98,** 9748–9753 (2001).

176. Compeau, P. E. C., Pevzner, P. A. & Tesler, G. How to apply de Bruijn graphs to genome assembly. *Nature Biotechnology* **29,** 987–991 (2011).

177. Schatz, M. C., Delcher, A. L. & Salzberg, S. L. Assembly of large genomes using second-generation sequencing. *Genome Research* **20,** 1165–1173 (2010).

178. Chau, T. T. *et al.* Antimicrobial drug resistance of *Salmonella enterica* serovar Typhi in Asia and molecular mechanism of reduced susceptibility to the fluoroquinolones. *Antimicrobial Agents and Chemotherapy* **51,** 4315–4323 (2007).

179. Parry, C. M. *et al.* The influence of reduced susceptibility to fluoroquinolones in *Salmonella enterica* serovar Typhi on the clinical response to ofloxacin therapy. *PLoS Neglected Tropical Diseases* **5,** e1163 (2011).

180. Maiden, M. C. J. Multilocus sequence typing of bacteria. *Annual Review of Microbiology* **60,** 561–588 (2006).

181. Maiden, M. C. J. *et al.* MLST revisited: the gene-by-gene approach to bacterial genomics. *Nature Reviews Microbiology* **11,** 728–736 (2013).

182. Zankari, E. *et al.* Identification of acquired antimicrobial resistance genes. *Journal of Antimicrobial Chemotherapy* **67,** 2640–2644 (2012).

183. Jia, B. *et al.* CARD 2017: expansion and model-centric curation of the comprehensive antibiotic resistance database. *Nucleic Acids Research* **45,** D566–D573 (2016).

184. Gupta, S. K. *et al.* ARG-ANNOT, a new bioinformatic tool to discover antibiotic resistance genes in bacterial genomes. *Antimicrobial Agents and Chemotherapy* **58,** 212–220 (2014).

185. Chen, L., Zheng, D., Liu, B., Yang, J. & Jin, Q. VFDB 2016: hierarchical and refined dataset

for big data analysis - 10 years on. *Nucleic Acids Research* **44,** D694–7 (2016).

186. Carattoli, A. *et al.* In silico detection and typing of plasmids using PlasmidFinder and plasmid multilocus sequence typing. *Antimicrobial Agents and Chemotherapy* **58,** 3895–3903 (2014).

187. Stoesser, N. *et al.* Evolutionary history of the global emergence of the *Escherichia coli* epidemic clone ST131. *mBio* **7,** e02162 (2016).

188. Holt, K. E. *et al.* Genomic analysis of diversity, population structure, virulence, and antimicrobial resistance in *Klebsiella pneumoniae*, an urgent threat to public health. *Proceedings of the National Academy of Sciences of the United States of America* **112,** (2015).

189. Leekitcharoenphon, P. *et al.* Global genomic epidemiology of *Salmonella* Typhimurium DT104. *Applied and Environmental Microbiology* **82,** 2516–2526 (2016).

190. Zhou, Z. *et al.* Transient Darwinian selection in *Salmonella enterica* serovar Paratyphi A during 450 years of global spread of enteric fever. *Proceedings of the National Academy of Sciences of the United States of America* **111,** 12199–12204 (2014).

191. Comas, I. *et al.* Out-of-Africa migration and Neolithic coexpansion of *Mycobacterium tuberculosis* with modern humans. *Nature Genetics* **45,** 1176–1182 (2013).

192. Harris, S. R. *et al.* Evolution of MRSA during hospital transmission and intercontinental spread. *Science* **327,** 469–474 (2010).

193. Zuckerkandl, E. & Pauling, L. in *Horizons in biochemistry* (eds. Kasha, M. & Pullman, B.) 189–225 (Academic Press, 1962).

194. Ochman, H. Neutral mutations and neutral substitutions in bacterial genomes. *Molecular Biology and Evolution* **20,** 2091–2096 (2003).

195. Rieux, A. & Balloux, F. Inferences from tip-calibrated phylogenies: a review and a practical guide. *Molecular Ecology* **25,** 1911–1924 (2016).

196. Achtman, M. Insights from genomic comparisons of genetically monomorphic bacterial pathogens. *Philosophical Transactions of the Royal Society B* **367,** 860–867 (2012).

197. Li, W. H., Tanimura, M. & Sharp, P. M. Rates and dates of divergence between AIDS virus

nucleotide sequences. *Molecular Biology and Evolution* **5,** 313–330 (1988).

198. Bromham, L. & Penny, D. The modern molecular clock. *Nature Reviews Genetics* **4,** 216–224 (2003).

199. Drummond, A., Forsberg, R. & Rodrigo, A. G. The inference of stepwise changes in substitution rates using serial sequence samples. *Molecular Biology and Evolution* **18,** 1365–1371 (2001).

200. Drummond, A. J. & Rambaut, A. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evolutionary Biology* **7,** 214 (2007).

201. Biek, R., Pybus, O. G., Lloyd-Smith, J. O. & Didelot, X. Measurably evolving pathogens in the genomic era. *Trends in Ecology & Evolution* **30,** 306–313 (2015).

202. Kotloff, K. L., Winickoff, J. P. & Ivanoff, B. Global burden of *Shigella* infections: implications for vaccine development and implementation of control strategies. *Bulletin of the World Health Organization* **77,** 651–666 (1999).

203. O'Loughlin, E. V. & Robins-Browne, R. M. Effect of Shiga toxin and Shiga-like toxins on eukaryotic cells. *Microbes and Infection* **3,** 493–507 (2001).

204. DuPont, H. L., Levine, M. M., Hornick, R. B. & Formal, S. B. Inoculum size in shigellosis and implications for expected mode of transmission. *Journal of Infectious Diseases* **159,** 1126–1128 (1989).

205. Gorden, J. & Small, P. L. Acid resistance in enteric bacteria. *Infection and Immunity* **61,** 364–367 (1993).

206. Islam, D. *et al.* Downregulation of bactericidal peptides in enteric infections: a novel immune escape mechanism with bacterial DNA as a potential regulator. *Nature Medicine* **7,** 180–185 (2001).

207. Wassef, J. S., Keren, D. F. & Mailloux, J. L. Role of M cells in initial antigen uptake and in ulcer formation in the rabbit intestinal loop model of shigellosis. *Infection and Immunity* **57,** 858–863 (1989).

208. Zychlinsky, A. *et al.* In vivo apoptosis in *Shigella flexneri* infections. *Infection and Immunity*

64, 5357–5365 (1996).

209. Zychlinsky, A., Prevost, M. C. & Sansonetti, P. J. *Shigella flexneri* induces apoptosis in infected macrophages. *Nature* **358,** 167–169 (1992).

210. Sansonetti, P. J., Ryter, A., Clerc, P., Maurelli, A. T. & Mounier, J. Multiplication of *Shigella flexneri* within HeLa cells: lysis of the phagocytic vacuole and plasmid-mediated contact hemolysis. *Infection and Immunity* **51,** 461–469 (1986).

211. Stevens, J. M., Galyov, E. E. & Stevens, M. P. Actin-dependent movement of bacterial pathogens. *Nature Reviews Microbiology* **4,** 91–101 (2006).

212. Schroeder, G. N. & Hilbi, H. Molecular pathogenesis of *Shigella spp.*: controlling host cell signaling, invasion, and death by type III secretion. *Clinical Microbiology Reviews* **21,** 134–156 (2008).

213. Cherla, R. P., Lee, S.-Y. & Tesh, V. L. Shiga toxins and apoptosis. *FEMS Microbiology Letters* **228,** 159–166 (2003).

214. Ewing, W. H. *Shigella* nomenclature. *Journal of Bacteriology* **57,** 633–638 (1949).

215. Marteyn, B., Gazi, A. & Sansonetti, P. *Shigella*: a model of virulence regulation in vivo. *Gut Microbes* **3,** 104–120 (2012).

216. Brenner, D. J., Fanning, G. R., Skerman, F. J. & Falkow, S. Polynucleotide sequence divergence among strains of *Escherichia coli* and closely related organisms. *Journal of Bacteriology* **109,** 953–965 (1972).

217. Ochman, H., Whittam, T. S., Caugant, D. A. & Selander, R. K. Enzyme polymorphism and genetic population structure in *Escherichia coli* and *Shigella*. *Journal of General Microbiology* **129,** 2715–2726 (1983).

218. Hartl, D. L. & Dykhuizen, D. E. The population genetics of *Escherichia coli*. *Annual Review of Genetics* **18,** 31–68 (1984).

219. Pupo, G. M., Karaolis, D. K., Lan, R. & Reeves, P. R. Evolutionary relationships among pathogenic and nonpathogenic *Escherichia coli* strains inferred from multilocus enzyme electrophoresis and mdh sequence studies. *Infection and Immunity* **65,** 2685–2692 (1997).

220. Rolland, K., Lambert-Zechovsky, N., Picard, B. & Denamur, E. *Shigella* and enteroinvasive

Escherichia coli strains are derived from distinct ancestral strains of *E. coli*. *Microbiology* **144,** 2667–2672 (1998).

221. Pupo, G. M., Lan, R. & Reeves, P. R. Multiple independent origins of *Shigella* clones of *Escherichia coli* and convergent evolution of many of their characteristics. *Proceedings of the National Academy of Sciences of the United States of America* **97,** 10567–10572 (2000).

222. Yang, J. *et al.* Revisiting the molecular evolutionary history of *Shigella spp.* *Journal of Molecular Evolution* **64,** 71–79 (2007).

223. The, H. C., Thanh, D. P., Holt, K. E., Thomson, N. R. & Baker, S. The genomic signatures of *Shigella* evolution, adaptation and geographical spread. *Nature Reviews Microbiology* **14,** 235–250 (2016).

224. Sansonetti, P. J., Kopecko, D. J. & Formal, S. B. Involvement of a plasmid in the invasive ability of *Shigella flexneri*. *Infection and Immunity* **35,** 852–860 (1982).

225. Parsot, C. *Shigella spp*. and enteroinvasive *Escherichia coli* pathogenicity factors. *FEMS Microbiology Letters* **252,** 11–18 (2005).

226. Tran Van Nhieu, G., Enninga, J., Sansonetti, P. & Grompone, G. Tyrosine kinase signaling and type III effectors orchestrating *Shigella* invasion. *Current Opinion in Microbiology* **8,** 16–20 (2005).

227. Lan, R., Lumb, B., Ryan, D. & Reeves, P. R. Molecular evolution of large virulence plasmid in *Shigella* clones and enteroinvasive *Escherichia coli*. *Infection and Immunity* **69,** 6303–6309 (2001).

228. Escobar-Páramo, P., Giudicelli, C., Parsot, C. & Denamur, E. The evolutionary history of *Shigella* and enteroinvasive *Escherichia coli* revised. *Journal of Molecular Evolution* **57,** 140–148 (2003).

229. Ingersoll, M., Groisman, E. A. & Zychlinsky, A. in *Pathogenicity islands and the evolution of pathogenic microbes* 49–65 (Springer Berlin Heidelberg, 2002).

230. Al-Hasani, K. *et al.* The *sigA* gene which is borne on the *she* pathogenicity island of *Shigella flexneri* 2a encodes an exported cytopathic protease involved in intestinal fluid accumulation. *Infection and Immunity* **68,** 2457–2463 (2000).

231. Moss, J. E., Cardozo, T. J., Zychlinsky, A. & Groisman, E. A. The *selC*-associated SHI-2

pathogenicity island of *Shigella flexneri*. *Molecular Microbiology* **33,** 74–83 (1999).

232. Vokes, S. A., Reeves, S. A., Torres, A. G. & Payne, S. M. The aerobactin iron transport system genes in *Shigella flexneri* are present within a pathogenicity island. *Molecular Microbiology* **33,** 63–73 (1999).

233. Peng, J. *et al.* The use of comparative genomic hybridization to characterize genome dynamics and diversity among the serotypes of *Shigella*. *BMC Genomics* **7,** 218 (2006).

234. Purdy, G. E. & Payne, S. M. The SHI-3 iron transport island of *Shigella boydii* 0-1392 carries the genes for aerobactin synthesis and transport. *Journal of Bacteriology* **183,** 4176–4182 (2001).

235. Huan, P. thi, Bastin, D. A., Whittle, B. L., Lindberg, A. A. & Verma, N. K. Molecular characterization of the genes involved in O-antigen modification, attachment, integration and excision in *Shigella flexneri* bacteriophage SfV. *Gene* **195,** 217–227 (1997).

236. Turner, S. A., Luck, S. N., Sakellaris, H., Rajakumar, K. & Adler, B. Nested deletions of the SRL pathogenicity island of *Shigella flexner*i 2a. *Journal of Bacteriology* **183,** 5535–5543 (2001).

237. Turner, S. A., Luck, S. N., Sakellaris, H., Rajakumar, K. & Adler, B. Molecular epidemiology of the SRL pathogenicity island. *Antimicrobial Agents and Chemotherapy* **47,** 727–734 (2003).

238. Turner, S. A., Luck, S. N., Sakellaris, H., Rajakumar, K. & Adler, B. Role of *attP* in integrase-mediated integration of the *Shigella* resistance locus pathogenicity island of *Shigella flexneri*. *Antimicrobial Agents and Chemotherapy* **48,** 1028–1031 (2004).

239. Mantis, N. J. & Sansonetti, P. J. The *nadB* gene of *Salmonella typhimurium* complements the nicotinic acid auxotrophy of *Shigella flexneri*. *Molecular & General Genetics* **252,** 626–629 (1996).

240. Prunier, A.-L., Schuch, R., Fernández, R. E. & Maurelli, A. T. Genetic structure of the *nadA* and *nadB* antivirulence loci in *Shigella spp. Journal of Bacteriology* **189,** 6482–6486 (2007).

241. Prunier, A.-L. *et al.* *nadA* and *nadB* of *Shigella flexneri* 5a are antivirulence loci responsible for the synthesis of quinolinate, a small molecule inhibitor of *Shigella* pathogenicity. *Microbiology* **153,** 2363–2372 (2007).

242. Maurelli, A. T., Fernández, R. E., Bloch, C. A., Rode, C. K. & Fasano, A. 'Black holes' and bacterial pathogenicity: a large genomic deletion that enhances the virulence of *Shigella*

spp. and enteroinvasive *Escherichia coli*. *Proceedings of the National Academy of Sciences of the United States of America* **95,** 3943–3948 (1998).

243. Barbagallo, M. *et al.* A new piece of the *Shigella* pathogenicity puzzle: spermidine accumulation by silencing of the *speG* gene. *PLoS ONE* **6,** e27226 (2011).

244. Nakata, N. *et al.* The absence of a surface protease, OmpT, determines the intercellular spreading ability of *Shigella*: the relationship between the *ompT* and *kcpA* loci. *Molecular Microbiology* **9,** 459–468 (1993).

245. Zhao, G. *et al.* A novel anti-virulence gene revealed by proteomic analysis in *Shigella flexneri* 2a. *Proteome Science* **8,** 30 (2010).

246. Ramos, H. C., Rumbo, M. & Sirard, J.-C. Bacterial flagellins: mediators of pathogenicity and host immune responses in mucosa. *Trends in Microbiology* **12,** 509–517 (2004).

247. Bergsten, G., Wullt, B. & Svanborg, C. *Escherichia coli*, fimbriae, bacterial persistence and host response induction in the human urinary tract. *International Journal of Medical Microbiology* **295,** 487–502 (2005).

248. Wei, J. *et al.* Complete genome sequence and comparative genomics of *Shigella flexneri* serotype 2a strain 2457T. *Infection and Immunity* **71,** 2775–2786 (2003).

249. Karaolis, D. K., Lan, R. & Reeves, P. R. Sequence variation in S*higella sonnei* (Sonnei), a pathogenic clone of *Escherichia coli*, over four continents and 41 years. *Journal of Clinical Microbiology* **32,** 796–802 (1994).

250. Holt, K. E. *et al. Shigella sonnei* genome sequencing and phylogenetic analysis indicate recent global dissemination from Europe. *Nature Genetics* **44,** 1056–1059 (2012).

251. Vinh, H. *et al.* A changing picture of shigellosis in southern Vietnam: shifting species dominance, antimicrobial susceptibility and clinical presentation. *BMC Infectious Diseases* **9,** 204 (2009).

252. Holt, K. E. *et al.* Tracking the establishment of local endemic populations of an emergent enteric pathogen. *Proceedings of the National Academy of Sciences of the United States of America* **110,** 17522–17527 (2013).

253. Sack, D. A., Hoque, A. T., Huq, A. & Etheridge, M. Is protection against shigellosis induced

by natural infection with *Plesiomonas shigelloides*? *Lancet* **343,** 1413–1415 (1994).

254. The, H. C. *et al.* The introduction and establishment of fluoroquinolone resistant Shigella sonnei into Bhutan. *Microbial Genomics* (2015).

255. Shiga, K. *Ueber den Erreger der dysenterie in Japan*. (Zentralbl Bakteriol Mikrobiol Hyg (Vorläufige Mitteilung), 1898).

256. Mata, L. J., Gangarosa, E. J., Cáceres, A., Perera, D. R. & Mejicanos, M. L. Epidemic Shiga bacillus dysentery in Central America. I. Etiologic investigations in Guatemala, 1969. *Journal of Infectious Diseases* **122,** 170–180 (1970).

257. Cobra, C. & Sack, D. A. The Control of Epidemic Dysentery in Africa: Overview, Recommendations, and Checklists. (1996).

258. Taylor, D. N. *et al.* Introduction and spread of multi-resistant *Shigella dysenteriae* I in Thailand. *The American Journal of Tropical Medicine and Hygiene* **40,** 77–85 (1989).

259. Rahaman, M. M., Khan, M. M., Aziz, K. M., Islam, M. S. & Kibriya, A. K. An outbreak of dysentery caused by *Shigella dysenteriae* type 1 on a coral island in the Bay of Bengal. *Journal of Infectious Diseases* **132,** 15–19 (1975).

260. Njamkepo, E. *et al.* Global phylogeography and evolutionary history of *Shigella dysenteriae* type 1. *Nature Microbiology* **1,** 16027 (2016).

261. Livio, S. *et al. Shigella* isolates from the global enteric multicenter study inform vaccine development. *Clinical Infectious Disease* **59,** 933–941 (2014).

262. Choi, S. Y. *et al.* Multilocus sequence typing analysis of *Shigella flexneri* isolates collected in Asian countries. *Journal of Medical Microbiology* **56,** 1460–1466 (2007).

263. Yang, Y. G., Song, M. K., Park, S. J. & Kim, S. W. Direct detection of *Shigella flexneri* and *Salmonella* Typhimurium in human feces by real-time PCR. *Journal of Microbiology and Biotechnology* **17,** 1616–1621 (2007).

264. Connor, T. R. *et al.* Species-wide whole genome sequencing reveals historical global spread and recent local persistence in *Shigella flexneri*. *eLife* **4,** e07335 (2015).

265. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler

transform. *Bioinformatics* **25,** 1754–1760 (2009).

266. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25,** 2078–2079 (2009).

267. Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10,** 421 (2009).

268. Jiang, C., Chen, C., Huang, Z., Liu, R. & Verdier, J. ITIS, a bioinformatics tool for accurate identification of transposon insertion sites using next-generation sequencing data. *BMC Bioinformatics* **16,** 72 (2015).

269. Adams, M. D., Bishop, B. & Wright, M. S. Quantitative assessment of insertion sequence impact on bacterial genome architecture. *Microbial Genomics* **2,** (2016).

270. Biswas, A., Gauthier, D. T., Ranjan, D. & Zubair, M. ISQuest: finding insertion sequences in prokaryotic sequence fragment data. *Bioinformatics* **31,** 3406–3412 (2015).

271. Gordon, N. C. *et al.* Prediction of *Staphylococcus aureus* antimicrobial resistance by whole-genome sequencing. *Journal of Clinical Microbiology* **52,** 1182–1191 (2014).

272. Kwong, J. C. *et al.* Prospective whole-genome sequencing enhances national surveillance of *Listeria monocytogenes*. *Journal of Clinical Microbiology* **54,** 333–342 (2016).

273. Deleo, F. R. *et al.* Molecular dissection of the evolution of carbapenem-resistant multilocus sequence type 258 *Klebsiella pneumoniae*. *Proceedings of the National Academy of Sciences of the United States of America* **111,** 4988–4993 (2014).

274. Hazen, T. H. *et al.* Comparative genomics of an IncA/C multidrug resistance plasmid from *Escherichia coli* and *Klebsiella* isolates from intensive care unit patients and the utility of whole-genome sequencing in health care settings. *Antimicrobial Agents and Chemotherapy* **58,** 4814–4825 (2014).

275. Conlan, S. *et al.* Single-molecule sequencing to track plasmid diversity of hospital-associated carbapenemase-producing Enterobacteriaceae. *Science Translational Medicine* **6,** 254ra126 (2014).

276. Wong, V. K. *et al.* Phylogeographical analysis of the dominant multidrug-resistant H58 clade of *Salmonella* Typhi identifies inter- and intracontinental transmission events. *Nature*

Genetics **47,** 632–639 (2015).

277. Lim, T. P. *et al.* Multiple genetic mutations associated with polymyxin resistance in *Acinetobacter baumannii*. *Antimicrobial Agents and Chemotherapy* **59,** 7899–7902 (2015).

278. CDC. *National Enteric Disease Surveillance:* Salmonella *Annual Report, 2012*. (2012).

279. Le Hello, S. *et al.* International spread of an epidemic population of *Salmonella enterica* serotype Kentucky ST198 resistant to ciprofloxacin. *Journal of Infectious Diseases* **204,** 675–684 (2011).

280. Johnson, T. J., Thorsness, J. L., Anderson, C. P. & Lynne, A. M. Horizontal gene transfer of a ColV plasmid has resulted in a dominant avian clonal type of *Salmonella enterica* serovar Kentucky. *PLoS ONE* **5,** e15524 (2010).

281. Foley, S. L. *et al.* Population dynamics of *Salmonella enterica* serotypes in commercial egg and poultry production. *Applied and Environmental Microbiology* **77,** 4273–4279 (2011).

282. Le Hello, S. *et al.* Early strains of multidrug-resistant *Salmonella enterica* serovar Kentucky sequence type 198 from Southeast Asia harbor *Salmonella* genomic island 1-J variants with a novel insertion sequence. *Antimicrobial Agents and Chemotherapy* **56,** 5096–5102 (2012).

283. Weill, F.-X. *et al.* Ciprofloxacin-resistant *Salmonella* Kentucky in travelers. *Emerging Infectious Diseases* **12,** 1611–1612 (2006).

284. Collard, J.-M. *et al.* Travel-acquired salmonellosis due to *Salmonella* Kentucky resistant to ciprofloxacin, ceftriaxone and co-trimoxazole and associated with treatment failure. *Journal of Antimicrobial Chemotherapy* **60,** 190–192 (2007).

285. Fricke, W. F. *et al.* Antimicrobial resistance-conferring plasmids with similarity to virulence plasmids from avian pathogenic *Escherichia coli* strains in *Salmonella enterica* serovar Kentucky isolates from poultry. *Applied and Environmental Microbiology* **75,** 5963–5971 (2009).

286. Le Hello, S. *et al.* The global establishment of a highly-fluoroquinolone resistant *Salmonella enterica* serotype Kentucky ST198 strain. *Frontiers in Microbiology* **4,** 395 (2013).

287. Sivaramalingam, T., Pearl, D. L., McEwen, S. A., Ojkic, D. & Guerin, M. T. A temporal study of *Salmonella* serovars from fluff samples from poultry breeder hatcheries in Ontario between

1998 and 2008. *The Canadian Journal of Veterinary Research* **77,** 12–23 (2013).

288. Mulvey, M. R. *et al.* Ciprofloxacin-resistant *Salmonella enterica* serovar Kentucky in Canada. *Emerging Infectious Diseases* **19,** 999–1001 (2013).

289. Boyd, D., Cloeckaert, A., Chaslus-Dancla, E. & Mulvey, M. R. Characterization of variant *Salmonella* genomic island 1 multidrug resistance regions from serovars Typhimurium DT104 and Agona. *Antimicrobial Agents and Chemotherapy* **46,** 1714–1722 (2002).

290. Hall, R. M. *Salmonella* genomic islands and antibiotic resistance in *Salmonella enterica*. *Future Microbiology* **5,** 1525–1538 (2010).

291. Doublet, B., Boyd, D., Mulvey, M. R. & Cloeckaert, A. The *Salmonella* genomic island 1 is an integrative mobilizable element. *Molecular Microbiology* **55,** 1911–1924 (2005).

292. Carraro, N., Matteau, D., Luo, P., Rodrigue, S. & Burrus, V. The master activator of IncA/C conjugative plasmids stimulates genomic islands and multidrug resistance dissemination. *PLoS Genetics* **10,** e1004714 (2014).

293. Douard, G., Praud, K., Cloeckaert, A. & Doublet, B. The *Salmonella* Genomic Island 1 is specifically mobilized in trans by the IncA/C multidrug resistance plasmid family. *PLoS ONE* **5,** e15302 (2010).

294. Kiss, J., Nagy, B. & Olasz, F. Stability, entrapment and variant formation of *Salmonella* genomic island 1. *PLoS ONE* **7,** e32497 (2012).

295. Levings, R. S., Djordjevic, S. P. & Hall, R. M. SGI2, a relative of *Salmonella* genomic island SGI1 with an independent origin. *Antimicrobial Agents and Chemotherapy* **52,** 2529–2537 (2008).

296. Levings, R. S., Partridge, S. R., Djordjevic, S. P. & Hall, R. M. SGI1-K, a variant of the SGI1 genomic island carrying a mercury resistance region, in *Salmonella enterica* serovar Kentucky. *Antimicrobial Agents and Chemotherapy* **51,** 317–323 (2007).

297. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9,** 357–359 (2012).

298. Croucher, N. J. *et al.* Rapid phylogenetic analysis of large samples of recombinant bacterial

whole genome sequences using Gubbins. *Nucleic Acids Research* **43,** gku1196–e15 (2014).

299. Drummond, A. J., Suchard, M. A., Xie, D. & Rambaut, A. Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Molecular Biology and Evolution* **29,** 1969–1973 (2012).

300. Rambaut, A., Suchard, M. A., Xie, D. & Drummond, A. J. Tracer v1.6. (2014).

301. United Nations Statistics Division. Composition of macro geographical (continental) regions, geographical sub-regions, and selected economic and other groupings. (2013). at <http://unstats.un.org/unsd/methods/m49/m49regin.htm>

302. Hannon Lab. FastXToolkit. (2010). at <http://hannonlab.cshl.edu/fastx_toolkit/>

303. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: A flexible trimmer for Illumina Sequence Data. *Bioinformatics* **30,** (2014).

304. Bankevich, A. *et al.* SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of Computational Biology* **19,** 455–477 (2012).

305. Boetzer, M., Henkel, C. V., Jansen, H. J., Butler, D. & Pirovano, W. Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics* **27,** 578–579 (2011).

306. Boetzer, M. & Pirovano, W. Toward almost closed genomes with GapFiller. *Genome Biology* **13,** 1 (2012).

307. Assefa, S., Keane, T. M., Otto, T. D., Newbold, C. & Berriman, M. ABACAS: algorithm-based automatic contiguation of assembled sequences. *Bioinformatics* **25,** 1968–1969 (2009).

308. Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30,** (2014).

309. Wick, R. R., Schultz, M. B., Zobel, J. & Holt, K. E. Bandage: interactive visualization of de novo genome assemblies. *Bioinformatics* **31,** 3350–3352 (2015).

310. Mollet, B., Iida, S., Shepherd, J. & Arber, W. Nucleotide sequence of IS*26*, a new prokaryotic mobile genetic element. *Nucleic Acids Research* **11,** 6319–6330 (1983).

311. Harmer, C. J., Moran, R. A. & Hall, R. M. Movement of IS*26*-associated antibiotic resistance genes occurs via a translocatable unit that includes a single IS*26* and preferentially inserts adjacent to another IS*26*. *mBio* **5,** e01801–14 (2014).

312. Parry, C. M., Hien, T. T., Dougan, G., White, N. J. & Farrar, J. J. Typhoid fever. *The New*

England Journal of Medicine **347,** 1770–1782 (2002).

313. Kothari, A., Pruthi, A. & Chugh, T. D. The burden of enteric fever. *Journal of Infection in Developing Countries* **2,** 253–259 (2008).

314. Crump, J. A. & Mintz, E. D. Global trends in typhoid and paratyphoid fever. *Clinical Infectious Disease* **50,** 241–246 (2010).

315. Mahon, B. E., Newton, A. E. & Mintz, E. D. Effectiveness of typhoid vaccination in US travelers. *Vaccine* **32,** 3577–3579 (2014).

316. Colquhoun, J. & Weetch, R. S. Resistance to chloramphenicol developing during treatment of typhoid fever. *Lancet* **2,** 621–623 (1950).

317. Mirza, S. H., Beeching, N. J. & Hart, C. A. Multi-drug resistant typhoid: a global problem. *Journal of Medical Microbiology* **44,** 317–319 (1996).

318. Hermans, P. W. *et al.* Molecular typing of *Salmonella typhi* strains from Dhaka (Bangladesh) and development of DNA probes identifying plasmid-encoded multidrug-resistant isolates. *Journal of Clinical Microbiology* **34,** 1373–1379 (1996).

319. Phan, M. D. *et al.* Variation in *Salmonella enterica* serovar Typhi IncHI1 plasmids during the global spread of resistant typhoid fever. *Antimicrobial Agents and Chemotherapy* **53,** 716–727 (2009).

320. Touchon, M. *et al.* Organised genome dynamics in the *Escherichia coli* species results in highly diverse adaptive paths. *PLoS Genetics* **5,** e1000344 (2009).

321. Kaye, K. S. & Pogue, J. M. Infections caused by resistant Gram-negative bacteria: epidemiology and management. *Pharmacotherapy* **35,** 949–962 (2015).

322. Rice, L. B. Federal funding for the study of antimicrobial resistance in nosocomial pathogens: no ESKAPE. *Journal of Infectious Diseases* **197,** 1079–1081 (2008).

323. Geisinger, E. & Isberg, R. R. Antibiotic modulation of capsular exopolysaccharide and virulence in *Acinetobacter baumannii*. *PLoS Pathogens* **11,** e1004691 (2015).

324. Bergogne-Bérézin, E. & Towner, K. J. *Acinetobacter spp.* as nosocomial pathogens: microbiological, clinical, and epidemiological features. *Clinical Microbiology Reviews* **9,**

148–165 (1996).

325. Hamidian, M. & Hall, R. M. Tn*6168,* a transposon carrying an ISAba1-activated *ampC* gene and conferring cephalosporin resistance in *Acinetobacter baumannii. Journal of Antimicrobial Chemotherapy* **69,** 77–80 (2014).

326. Javan, A. O., Shokouhi, S. & Sahraei, Z. A review on colistin nephrotoxicity. *European Journal of Clinical Pharmacology* **71,** 801–810 (2015).

327. Nhu, N. T. K. *et al.* Emergence of carbapenem-resistant *Acinetobacter baumannii* as the major cause of ventilator-associated pneumonia in intensive care unit patients at an infectious disease hospital in southern Vietnam. *Journal of Medical Microbiology* **63,** 1386–1394 (2014).

328. Newton, B. A. The properties and mode of action of the polymyxins. *Bacteriological Reviews* **20,** 14–27 (1956).

329. Gunn, J. S. *et al.* PmrA-PmrB-regulated genes necessary for 4-aminoarabinose lipid A modification and polymyxin resistance. *Molecular Microbiology* **27,** 1171–1182 (1998).

330. Macfarlane, E. L., Kwasnicka, A., Ochs, M. M. & Hancock, R. E. PhoP-PhoQ homologues in *Pseudomonas aeruginosa* regulate expression of the outer-membrane protein OprH and polymyxin B resistance. *Molecular Microbiology* **34,** 305–316 (1999).

331. Moffatt, J. H. *et al.* Colistin resistance in *Acinetobacter baumannii* is mediated by complete loss of lipopolysaccharide production. *Antimicrobial Agents and Chemotherapy* **54,** 4971–4977 (2010).

332. Liu, Y.-Y. *et al.* Emergence of plasmid-mediated colistin resistance mechanism MCR-1 in animals and human beings in China: a microbiological and molecular biological study. *Lancet Infectious Diseases* **16,** 161–168 (2016).

333. Qureshi, Z. A. *et al.* Colistin-resistant *Acinetobacter baumannii*: beyond carbapenem resistance. *Clinical Infectious Disease* **60,** 1295–1303 (2015).

334. Kasiakou, S. K. *et al.* Combination therapy with intravenous colistin for management of infections due to multidrug-resistant Gram-negative bacteria in patients without cystic fibrosis. *Antimicrobial Agents and Chemotherapy* **49,** 3136–3146 (2005).

335. Hornsey, M. *et al.* Whole-genome comparison of two *Acinetobacter baumannii* isolates from a single patient, where resistance developed during tigecycline therapy. *Journal of*

Antimicrobial Chemotherapy **66,** 1499–1503 (2011).

336. Young, D. M. & Ornston, L. N. Functions of the mismatch repair gene *mutS* from *Acinetobacter sp.* strain ADP1. *Journal of Bacteriology* **183,** 6822–6831 (2001).

337. Malinverni, J. C. & Silhavy, T. J. An ABC transport system that maintains lipid asymmetry in the Gram-negative outer membrane. *Proceedings of the National Academy of Sciences of the United States of America* **106,** 8009–8014 (2009).

338. Köser, C. U. *et al.* Routine use of microbial whole genome sequencing in diagnostic and public health microbiology. *PLoS Pathogens* **8,** e1002824 (2012).

339. Azarian, T. *et al.* Whole-genome sequencing for outbreak investigations of methicillin-resistant *Staphylococcus aureus* in the neonatal intensive care unit: time for routine practice? *Infection Control and Hospital Epidemiology* **36,** 777–785 (2015).

340. Loman, N. J. & Pallen, M. J. Twenty years of bacterial genome sequencing. *Nature Reviews Microbiology* **13,** 787–794 (2015).

341. Yang, F. *et al.* Genome dynamics and diversity of *Shigella* species, the etiologic agents of bacillary dysentery. *Nucleic Acids Research* **33,** 6445–6458 (2005).

342. Varani, A. M., Siguier, P., Gourbeyre, E., Charneau, V. & Chandler, M. ISsaga is an ensemble of web-based methods for high throughput identification and semi-automatic annotation of insertion sequences in prokaryotic genomes. *Genome Biology* **12,** R30 (2011).

343. Otto, T. D., Dillon, G. P., Degrave, W. S. & Berriman, M. RATT: Rapid Annotation Transfer Tool. *Nucleic Acids Research* **39,** gkq1268–e57 (2011).

344. Overbeek, R. *et al.* The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). *Nucleic Acids Research* **42,** D206–14 (2014).

345. Aziz, R. K. *et al.* The RAST Server: Rapid Annotations using Subsystems Technology. *BMC Genomics* **9,** 75 (2008).

346. Rice, P., Longden, I. & Bleasby, A. EMBOSS: the European Molecular Biology Open Software Suite. *Trends in Genetics* **16,** 276–277 (2000).

347. Schliep, K. P. phangorn: phylogenetic analysis in R. *Bioinformatics* **27,** 592–593 (2011).

348. Kurtz, S., Phillippy, A. & Delcher, A. L. Versatile and open software for comparing large

genomes. *Genome Biology* **5,** (2004).

349. Eduardo P. C. Rocha. The replication-related organization of bacterial genomes. *Microbiology* **150,** 1609–1627 (2004).

350. Li, G., Laturnus, C., Ewers, C. & Wieler, L. H. Identification of genes required for avian *Escherichia coli* septicemia by signature-tagged mutagenesis. *Infection and Immunity* **73,** 2818–2827 (2005).

351. Stratmann, T., Madhusudan, S. & Schnetz, K. Regulation of the *yjjQ-bglJ* operon, encoding LuxR-type transcription factors, and the divergent *yjjP* gene by H-NS and LeuO. *Journal of Bacteriology* **190,** 926–935 (2008).

352. Wiebe, H. *et al. YjjQ* represses transcription of *flhDC* and additional loci in *Escherichia coli*. *Journal of Bacteriology* **197,** 2713–2720 (2015).

353. Wang, C., Yang, L., Shah, A. A., Choi, E.-S. & Kim, S.-W. Dynamic interplay of multidrug transporters with TolC for isoprenol tolerance in *Escherichia coli*. *Scientific Reports* **5,** 16505 (2015).

354. Cascales, E. *et al.* Colicin Biology. *Microbiology and Molecular Biology Reviews* **71,** 158–229 (2007).

355. Calcuttawala, F., Hariharan, C., Pazhani, G. P., Ghosh, S. & Ramamurthy, T. Activity spectrum of colicins produced by *Shigella sonnei* and genetic mechanism of colicin resistance in conspecific *S. sonnei* strains and *Escherichia coli*. *Antimicrobial Agents and Chemotherapy* **59,** 152–158 (2015).

356. Schmidt, O. *et al. prlF* and *yhaV* encode a new toxin-antitoxin system in *Escherichia coli*. *Journal of Molecular Biology* **372,** 894–905 (2007).

357. Chosa, H. *et al.* Loss of virulence in *Shigella* strains preserved in culture collections due to molecular alteration of the invasion plasmid. *Microbial Pathogenesis* **6,** 337–342 (1989).

358. Feng, Y., Chen, Z. & Liu, S.-L. Gene decay in *Shigella* as an incipient stage of host-adaptation. *PLoS ONE* **6,** e27754 (2011).

359. Bravo, V., Puhar, A., Sansonetti, P., Parsot, C. & Toro, C. S. Distinct mutations led to

inactivation of type 1 fimbriae expression in *Shigella spp*. *PLoS ONE* **10,** e0121785 (2015).

360. Masaki, H., Yajima, S., Akutsu-Koide, A., Ohta, T. & Uozumi, T. in *Bacteriocins, microcins and lantibiotics* 379–395 (Springer Berlin Heidelberg, 1992).

361. Kania, D. A. *et al.* Genome diversity of *Shigella boydii*. *Pathogens and Disease* **74,** ftw027 (2016).

362. Zhou, Y., Liang, Y., Lynch, K. H., Dennis, J. J. & Wishart, D. S. PHAST: A Fast Phage Search Tool. *Nucleic Acids Research* **39,** W347–W352 (2011).

363. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30,** 1312–1313 (2014).

364. Wall, D. P., Fraser, H. B. & Hirsh, A. E. Detecting putative orthologs. *Bioinformatics* **19,** 1710–1711 (2003).

365. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* (2004).

366. Ingle, D. J. *et al.* In silico serotyping of *E. coli* from short read data identifies limited novel O-loci but extensive diversity of O:H serotype combinations within and between pathogenic lineages. *Microbial Genomics* **2,** (2016).

367. Rasko, D. A. *et al.* The pangenome structure of Escherichia coli: comparative genomic analysis of E. coli commensal and pathogenic isolates. *Journal of Bacteriology* **190,** 6881–6893 (2008).

368. Inouye, M. *et al.* SRST2: Rapid genomic surveillance for public health and hospital microbiology labs. *Genome Medicine* **6,** 90 (2014).

369. Toh, H. *et al.* Complete genome sequence of the wild-type commensal *Escherichia coli* strain SE15, belonging to phylogenetic group B2. *Journal of Bacteriology* **192,** 1165–1166 (2010).

370. Petty, N. K. *et al.* Global dissemination of a multidrug resistant *Escherichia coli* clone. *Proceedings of the National Academy of Sciences of the United States of America* **111,** 5694–5699 (2014).

371. Price, L. B. *et al.* The epidemic of extended-spectrum-β-lactamase-producing *Escherichia*

coli ST131 is driven by a single highly pathogenic subclone, H30-Rx. *mBio* **4,** e00377–13 (2013).

372. Latif, H., Li, H. J., Charusanti, P., Palsson, B. Ø. & Aziz, R. K. A gapless, unambiguous genome sequence of the enterohemorrhagic *Escherichia coli* O157:H7 strain EDL933. *Genome Announcements* **2,** e00821–14–e00821–14 (2014).

373. Rohde, H. *et al.* Open-Source Genomic Analysis of Shiga-ToxinProducing *E. coli* O104:H4. *The New England Journal of Medicine* **365,** 718–724 (2011).

374. Escoubas, J. M., Lane, D. & Chandler, M. Is the IS*1* transposase, InsAB', the only IS*1*-encoded protein required for efficient transposition? *Journal of Bacteriology* **176,** 5864–5867 (1994).

375. Escoubas, J. M. *et al.* Translational control of transposition activity of the bacterial insertion sequence IS*1*. *The EMBO Journal* **10,** 705–712 (1991).

376. Ashida, H., Toyotome, T., Nagai, T. & Sasakawa, C. *Shigella* chromosomal IpaH proteins are secreted via the type III secretion system and act as effectors. *Molecular Microbiology* **63,** 680–693 (2007).

377. Liu, J. Y., Miller, P. F., Gosink, M. & Olson, E. R. The identification of a new family of sugar efflux pumps in Escherichia coli. *Molecular Microbiology* **31,** 1845–1851 (1999).

378. Snider, J. *et al.* Formation of a Distinctive Complex between the Inducible Bacterial Lysine Decarboxylase and a Novel AAA+ ATPase. *Journal of Biological Chemistry* **281,** 1532–1546 (2006).

379. Niki, H., Jaffé, A., Imamura, R., Ogura, T. & Hiraga, S. The new gene *mukB* codes for a 177 kd protein with coiled-coil domains involved in chromosome partitioning of *E. coli*. *The EMBO Journal* **10,** 183–193 (1991).

380. Yamanaka, K., Ogura, T., Niki, H. & Hiraga, S. Identification and characterization of the *smbA* gene, a suppressor of the *mukB* null mutant of *Escherichia coli*. *Journal of Bacteriology* **174,** 7517–7526 (1992).

381. Yamanaka, K., Mitani, T., Ogura, T., Niki, H. & Hiraga, S. Cloning, sequencing, and characterization of multicopy suppressors of a *mukB* mutation in *Escherichia coli*. *Molecular Microbiology* **13,** 301–312 (1994).

382. Hu, K. H. *et al.* Overproduction of three genes leads to camphor resistance and

chromosome condensation in *Escherichia coli*. *Genetics* **143,** 1521–1532 (1996).

383. Adachi, S. & Hiraga, S. Mutants suppressing novobiocin hypersensitivity of a *mukB* null mutation. *Journal of Bacteriology* **185,** 3690–3695 (2003).

384. Kido, M. *et al.* RNase E polypeptides lacking a carboxyl-terminal half suppress a *mukB* mutation in *Escherichia coli*. *Journal of Bacteriology* **178,** 3917–3925 (1996).

385. Livio, S. *et al.* *Shigella* isolates from the global enteric multicenter study inform vaccine development. *Clinical Infectious Disease* **59,** 933–941 (2014).

386. Devoid, S. *et al.* Automated genome annotation and metabolic model reconstruction in the SEED and Model SEED. *Methods in Molecular Biology* **985,** 17–45 (2013).

387. Hershberg, R., Tang, H. & Petrov, D. A. Reduced selection leads to accelerated gene loss in *Shigella*. *Genome Biology* **8,** R164 (2007).

388. Barrick, J. E. *et al.* Identifying structural variation in haploid microbial genomes from short-read resequencing data using breseq. *BMC Genomics* **15,** 1039 (2014).

389. Thung, D. T. *et al.* Mobster: accurate detection of mobile element insertions in next generation sequencing data. *Genome Biology* **15,** 488 (2014).

390. Nakagome, M. *et al.* Transposon Insertion Finder (TIF): a novel program for detection of de novo transpositions of transposable elements. *BMC Bioinformatics* **15,** 71 (2014).

391. Platzer, A., Nizhynska, V. & Long, Q. TE-Locate: A tool to locate and group transposable element occurrences using paired-end next-generation sequencing data. *Biology* **1,** 395–410 (2012).

392. Hormozdiari, F. *et al.* Next-generation VariationHunter: combinatorial algorithms for transposon insertion discovery. *Bioinformatics* **26,** i350–7 (2010).

393. Lee, K. Y., Hopkins, J. D. & Syvanen, M. Direct involvement of IS*26* in an antibiotic resistance operon. *Journal of Bacteriology* **172,** 3229–3236 (1990).

394. Stephens, Z. D. *et al.* Big Data: Astronomical or Genomical? *PLoS Biology* **13,** e1002195 (2015).

395. Allard, M. W. *et al.* Practical value of food pathogen traceability through building a whole-genome sequencing network and database. *Journal of Clinical Microbiology* **54,**

1975–1983 (2016).

396. Hazen, T. H. *et al.* Investigating the relatedness of enteroinvasive *Escherichia coli* to other *E. coli* and *Shigella* isolates by using comparative genomics. *Infection and Immunity* **84,** 2362–2371 (2016).

397. Saito, T., Chibazakura, T., Takahashi, K., Yoshikawa, H. & Sekine, Y. Measurements of transposition frequency of insertion sequence IS*1* by GFP hop-on assay. *The Journal of General and Applied Microbiology* **56,** 187–192 (2010).

398. Freed, N. E., Bumann, D. & Silander, O. K. Combining *Shigella* Tn-seq data with gold-standard *E. coli* gene deletion data suggests rare transitions between essential and non-essential gene functionality. *BMC Microbiology* **16,** 203 (2016).

399. Monk, J. M. *et al.* Genome-scale metabolic reconstructions of multiple *Escherichia coli* strains highlight strain-specific adaptations to nutritional environments. *Proceedings of the National Academy of Sciences of the United States of America* **110,** 20338–20343 (2013).

400. Cole, S. T. *et al.* Massive gene decay in the leprosy bacillus. *Nature* **409,** 1007–1011 (2001).

401. Bulach, D. M. *et al.* Genome reduction in *Leptospira borgpetersenii* reflects limited transmission potential. *Proceedings of the National Academy of Sciences of the United States of America* **103,** 14560–14565 (2006).

402. Doig, K. D. *et al.* On the origin of *Mycobacterium ulcerans*, the causative agent of Buruli ulcer. *BMC Genomics* **13,** 258 (2012).

403. Leavis, H. L. *et al.* Insertion sequencedriven diversification creates a globally dispersed emerging multiresistant subspecies of *E. faecium*. *PLoS Pathogens* **3,** e7 (2007).

Appendices

Supplmentary Table 1: Genes with conserved inactivatios in >90% of all *S. sonnei* genomes, or conserved inactivation in >90% of *S. sonnei* genomes from a particular lineage

Gene	Product	Inactivations	% Lineage I	% Lineage II	% Lineage III
ALL GENOMES					
SSON53G_RS00210	hypothetical protein	1	87.5	100	98.70
SSON53G_RS00750	adhesin	3	43.75	96.97	98.70
SSON53G_RS00790	fimbrial assembly protein	3	87.5	100	100
SSON53G_RS01375	hypothetical protein	3	75	100	94.81
SSON53G_RS02180	sel1 repeat family protein	1	100	96.97	98.70
SSON53G_RS02565	sugar ABC transporter permease	1	100	100	100
SSON53G_RS03305	hypothetical protein	1	87.5	90.91	96.10
SSON53G_RS03405	hypothetical protein	1	93.75	81.82	97.40
SSON53G_RS04290	TonB-dependent receptor	1	81.25	96.97	97.40
SSON53G_RS04835	LysR family transcriptional regulator	1	100	93.94	97.40
SSON53G_RS04915	competence protein ComEC	3	100	96.97	98.70
SSON53G_RS05465	aminoacrylate peracid reductase	3	100	100	98.70
SSON53G_RS07150	hypothetical protein	3	87.5	100	98.70
SSON53G_RS07840	membrane protein	1	100	84.85	98.70
SSON53G_RS10585	dienelactone hydrolase	3	75	96.97	98.70
SSON53G_RS11645	hypothetical protein	3	93.75	100	98.70
SSON53G_RS11680	transcriptional regulator	3	37.5	100	97.40
SSON53G_RS12530	transporter	1	93.75	90.91	92.21
SSON53G_RS13205	histidine kinase	3	100	96.97	98.70
SSON53G_RS13255	hypothetical protein	3	93.75	96.97	98.70
SSON53G_RS15120	membrane protein	3	100	100	100
SSON53G_RS16045	tail protein	3	37.5	100	98.70
SSON53G_RS16805	serine/threonine protein phosphatase	1	100	100	98.70
SSON53G_RS16815	3-hydroxyisobutyrate dehydrogenase	1	100	100	98.70
SSON53G_RS17070	hypothetical protein	3	100	96.97	100
SSON53G_RS17460	hypothetical protein	3	100	100	100
SSON53G_RS17815	LysR family transcriptional regulator	3	100	100	100

Supplmentary Table 1: Genes with conserved inactivatios in >90% of all *S. sonnei* genomes, or conserved inactivation in >90% of *S. sonnei* genomes from a particular lineage

Gene	Product	Inactivations	% Lineage I	% Lineage II	% Lineage III
SSON53G_RS20465	hypothetical protein	3	100	90.91	100
SSON53G_RS20485	membrane protein	3	100	100	100
SSON53G_RS20545	ATP-dependent DNA helicase RecQ	3	100	90.91	93.51
SSON53G_RS21290	leader peptide IlvB	1	93.75	100	98.70
SSON53G_RS21545	transporter	1	93.75	93.94	100
SSON53G_RS21660	DNA repair protein RadC	1	93.75	96.97	98.70
SSON53G_RS22550	ATP-dependent protease	1	93.75	100	100
SSON53G_RS22820	hypothetical protein	1	87.5	100	97.40
SSON53G_RS23185	sugar kinase	1	87.5	90.91	98.70
SSON53G_RS24525	acetyl-CoA synthetase	1	87.5	87.88	93.51
SSON53G_RS25050	transporter	1	100	100	98.70
SSON53G_RS25525	DNA-binding protein	3	93.75	100	100
SSON53G_RS02720	bacteriophage N4 adsorption protein B	3	93.75	100	98.70
SSON53G_RS02730	type IV secretion protein Rhs	3	93.75	100	98.70
SSON53G_RS02735	hypothetical protein	3	93.75	100	98.70
SSON53G_RS20575	acetyltransferase	3	100	100	98.70
SSON53G_RS00405	sugar MFS transporter	1	93.75	100	100
SSON53G_RS06275	DNA polymerase V subunit UmuC	1	87.5	87.88	96.10
SSON53G_RS06600	flagellar biosynthesis protein FlhB	1	93.75	100	100
SSON53G_RS10650	membrane protein	1	93.75	90.91	97.40
SSON53G_RS10660	sugar ABC transporter ATP-binding protein	1	93.75	90.91	97.40
SSON53G_RS15645	transposase	1	87.5	96.97	90.91
SSON53G_RS15840	L-aspartate oxidase	1	100	100	100
SSON53G_RS16335	membrane protein	1	100	100	100
SSON53G_RS18025	hypothetical protein	1	87.5	96.97	94.81
SSON53G_RS18535	hypothetical protein	1	100	100	98.70
SSON53G_RS22935	hypothetical protein	1	100	100	100
SSON53G_RS24330	hypothetical protein	1	100	100	100
SSON53G_RS24340	hypothetical protein	1	100	100	100
SSON53G_RS24785	hypothetical protein	1	93.75	93.94	97.40
SSON53G_RS25855	penicillin acylase	1	100	93.94	98.70
SSON53G_RS25860	hypothetical protein	1	100	93.94	98.70

Supplmentary Table 1: Genes with conserved inactivatios in >90% of all *S. sonnei* genomes, or conserved inactivation in >90% of *S. sonnei* genomes from a particular lineage

Gene	Product	Inactivations	% Lineage I	% Lineage II	% Lineage III
LINEAGE I					
SSON53G_RS01550	hypothetical protein	3	100	87.88	87.01
SSON53G_RS02180	sel1 repeat family protein	1	100	96.97	98.70
SSON53G_RS04835	LysR family transcriptional regulator	1	100	93.94	97.40
SSON53G_RS04915	competence protein ComEC	3	100	96.97	98.70
SSON53G_RS06970	hypothetical protein	1	100	0	6.49
SSON53G_RS07840	membrane protein	1	100	84.85	98.70
SSON53G_RS13205	histidine kinase	3	100	96.97	98.70
SSON53G_RS13270	adhesin	1	100	0	1.30
SSON53G_RS13500	hypothetical protein	1	100	0	0
SSON53G_RS18880	hypothetical protein	1	100	0	0
SSON53G_RS20545	ATP-dependent DNA helicase RecQ	3	100	90.91	93.51
SSON53G_RS25230	PTS ascorbate transporter subunit IIA	1	100	36.36	98.70
SSON53G_RS25900	Pyoverdin chromophore biosynthetic protein pvcC	1	100	3.03	0
SSON53G_RS22945	carboxymethylenebut enolidase	1	100	0	0
SSON53G_RS18050	membrane protein	1	100	96.97	0
SSON53G_RS25855	penicillin acylase	1	100	93.94	98.70
SSON53G_RS25860	hypothetical protein	1	100	93.94	98.70
LINEAGE II					
SSON53G_RS00210	hypothetical protein	1	87.50	100	98.70
SSON53G_RS01375	hypothetical protein	3	75.00	100	94.81
SSON53G_RS05465	aminoacrylate peracid reductase	3	100	100	98.70
SSON53G_RS07150	hypothetical protein	3	87.50	100	98.70
SSON53G_RS08985	fimbrial chaperone protein FimC	1	25.00	100	0
SSON53G_RS11645	hypothetical protein	3	93.75	100	98.70
SSON53G_RS11680	transcriptional regulator	3	37.50	100	97.40
SSON53G_RS16045	tail protein	3	37.50	100	98.70

Supplmentary Table 1: Genes with conserved inactivatios in >90% of all *S. sonnei* genomes, or conserved inactivation in >90% of *S. sonnei* genomes from a particular lineage

Gene	Product	Inactivations	% Lineage I	% Lineage II	% Lineage III
SSON53G_RS16805	serine/threonine protein phosphatase	1	100	100	98.70
SSON53G_RS16815	3-hydroxyisobutyrate dehydrogenase	1	100.00	100	98.70
SSON53G_RS21290	leader peptide IlvB	1	93.75	100	98.70
SSON53G_RS22820	hypothetical protein	1	87.50	100	97.40
SSON53G_RS25050	transporter	1	100	100	98.70
SSON53G_RS02720	bacteriophage N4 adsorption protein B	3	93.75	100	98.70
SSON53G_RS02730	type IV secretion protein Rhs	3	93.75	100	98.70
SSON53G_RS02735	hypothetical protein	3	93.75	100	98.70
SSON53G_RS20575	acetyltransferase	3	100	100	98.70
SSON53G_RS00800	fimbrial assembly protein	1	25.00	100	98.70
SSON53G_RS04470	c-di-GMP phosphodiesterase	1	0	100	81.82
SSON53G_RS12565	molybdenum metabolism regulator	1	12.50	100	98.70
SSON53G_RS18035	membrane protein	1	93.75	100	0
SSON53G_RS18060	hypothetical protein	1	93.75	100	0
SSON53G_RS18535	hypothetical protein	1	100	100	98.70
SSON53G_RS19005	toxin YhaV	1	0	100	5.19
SSON53G_RS19975	general secretion pathway protein GspA	1	0	100	0
SSON53G_RS21555	permease	1	0	100	98.70
SSON53G_RS21560	alpha-glucosidase	1	0	100	98.70
LINEAGE III					
SSON53G_RS00790	fimbrial assembly protein	3	87.50	100	100
SSON53G_RS08045	enterotoxin	3	0	3.03	100
SSON53G_RS11530	protein sirB1	1	0	3.03	100
SSON53G_RS16960	CRISPR-associated protein CasA	1	0	0	100
SSON53G_RS17070	hypothetical protein	3	100	96.97	100
SSON53G_RS18150	DNA helicase	3	100	6.06	100
SSON53G_RS18525	fimbrial outer membrane usher protein StdB	3	6.25	6.06	100
SSON53G_RS20465	hypothetical protein	3	100	90.91	100

Supplmentary Table 1: Genes with conserved inactivatios in >90% of all *S. sonnei* genomes, or conserved inactivation in >90% of *S. sonnei* genomes from a particular lineage

Gene	Product	Inactivations	% Lineage I	% Lineage II	% Lineage III
SSON53G_RS21535	integrase	1	93.75	3.03	100
SSON53G_RS21545	transporter	1	93.75	93.94	100
SSON53G_RS22550	ATP-dependent protease	1	93.75	100	100
SSON53G_RS24875	lysine decarboxylase LdcC	3	0	100	100
SSON53G_RS25520	toxin-antitoxin biofilm protein TabA	3	0	100	100
SSON53G_RS25525	DNA-binding protein	3	93.75	100	100
SSON53G_RS00590	type IV pilin biogenesis protein	1	0	0	100
SSON53G_RS02600	membrane protein	1	0	0	100
SSON53G_RS02935	hypothetical protein	3	0	0	100
SSON53G_RS04830	inner membrane transporter YcaM	3	6.25	0	100
SSON53G_RS13465	protein ElaA	1	0	0	100
SSON53G_RS14555	PTS N-acetylmuramic acid transporter subunit IIBC	1	0	0	100
SSON53G_RS14615	ethanolamine ammonia lyase large subunit	1	0	0	100
SSON53G_RS17425	racemase	1	0	0	100
SSON53G_RS18615	transcriptional activator TtdR	1	0	0	100
SSON53G_RS20270	DNA utilization protein HofN	1	0	0	100
SSON53G_RS23330	hypothetical protein	1	0	0	100
SSON53G_RS23340	HTH-type transcriptional activator RhaR	1	0	0	100
SSON53G_RS24150	isocitrate lyase	1	0	0	100
SSON53G_RS00405	sugar MFS transporter	1	93.75	100	100
SSON53G_RS06600	flagellar biosynthesis protein FlhB	1	93.75	100	100
SSON53G_RS06660	trimethylamine N-oxide reductase I catalytic subunit	1	0	0	100
SSON53G_RS11320	hypothetical protein	1	0	9.09	100
SSON53G_RS13285	membrane protein	1	0	96.97	100
SSON53G_RS13660	transposase	1	68.75	0	100

Supplmentary Table 1: Genes with conserved inactivatios in >90% of all *S. sonnei* genomes, or conserved inactivation in >90% of *S. sonnei* genomes from a particular lineage

Gene	Product	Inactivations	% Lineage I	% Lineage II	% Lineage III
SSON53G_RS18250	ATP-binding protein type I deoxyribonuclease	1	0	0	100
SSON53G_RS19000	HsdR	1	0	0	100
SSON53G_RS26020	LuxR family transcriptional regulator	1	6.25	0	100

Supplmentary Table 2: Genes under balancing or negative selection in *S. sonnei*

Locus tag	Product	Gene	IS sites	Total interruptions	SEED	RAST Category	Tips affected	Degredation Index	% Lineage I	% Lineage II	% Lineage III
SSON53G_RS00750	adhesin	-	1	2	-	-	115	0.0087	43.75	96.97	98.70
SSON53G_RS01375	hypothetical protein	-	1	2	-	-	124	0.0726	75.00	100	94.81
SSON53G_RS01550	hypothetical protein	-	2	3	fig|216599.13.peg.258	-	118	0.0254	100	87.88	87.01
SSON53G_RS02685	transcriptional regulator	*fimZ*	2	2	fig|216599.13.peg.456	-	75	0.0133	0	0	97.40
SSON53G_RS04090	tail protein	-	7	7	fig|216599.13.peg.674	Phages, Prophages, Transposable elements, Plasmids	38	0.2105	12.50	24.24	36.36
SSON53G_RS07150	hypothetical protein	-	1	2	-	-	123	0.0081	87.50	100	98.70
SSON53G_RS08475	transporter	*ydiM*	3	4	fig|216599.13.peg.1399	-	62	0	12.50	0	77.92
SSON53G_RS09045	DUF4186 domain-containing protein	*yneG*	1	1	fig|216599.13.peg.1484	-	72	0	0	0	93.51
SSON53G_RS11265	sugar acetyltransfer-ase inhibitor	-	1	3	-	-	114	0.0088	81.25	66.67	96.10
SSON53G_RS11320	hypothetical protein	-	3	3	fig|216599.13.peg.1806	-	80	0.0000	0.00	9.09	100.00

Supplmentary Table 2: Genes under balancing or negative selection in *S. sonnei*

Locus tag	Product	Gene	IS sites	Total interruptions	SEED	RAST Category	Tips affected	Degredation Index	% Lineage I	% Lineage II	% Lineage III
SSON53G_RS11645	hypothetical protein	-	2	5	fig|216599.13.peg.1863	-	124	0.2419	93.75	100.00	98.70
SSON53G_RS11680	transcriptional regulator	-	2	4	-	-	120	0.2583	37.50	100.00	97.40
SSON53G_RS12450	PTS galactitol transporter subunit IIA	-	1	2	fig|216599.13.peg.1999	Carbohydrates	73	0.0137	0	0	94.81
SSON53G_RS13315	hypothetical protein	*yfaH*	1	1	fig|216599.13.peg.2141	Regulation and Cell signaling	70	0	0	0	90.91
SSON53G_RS13660	transposase	-	2	2	fig|216599.13.peg.2205	-	88	0	68.75	0	100.00
SSON53G_RS15520	tail protein	-	4	4	-	-	34	0.3529	25.00	24.24	28.57
SSON53G_RS15645	transposase	ISEc8	1	1	fig|216599.13.peg.2538	-	116	0	87.50	96.97	90.91
SSON53G_RS16065	tail protein	-	6	6	fig|216599.13.peg.2593	-	33	0.0909	25.00	21.21	28.57
SSON53G_RS17815	LysR family transcriptional regulator	*ygfI*	1	4	-	-	126	0.0079	100	100	100
SSON53G_RS18025	hypothetical protein	*yggM*	1	1	-	-	119	0.0000	87.50	96.97	94.81
SSON53G_RS18050	membrane protein	*ygiK*	2	2	-	-	48	0.0208	100	96.97	0.00
SSON53G_RS18535	hypothetical protein	*yqiI*	2	2	fig|216599.13.peg.2993	-	125	0.0240	100	100	98.70
SSON53G_RS18830	fimbrial protein	-	1	1	fig|216599.13.peg.3044	Membrane Transport	73	0	0	0	94.81

Supplmentary Table 2: Genes under balancing or negative selection in *S. sonnei*

Locus tag	Product	Gene	IS sites	Total interruptions	SEED	RAST Category	Tips affected	Degredation Index	% Lineage I	% Lineage II	% Lineage III
SSON53G_RS19000	type I deoxyribonuclease HsdR	*ygfI*	1	1	fig|216599.13.peg.3074	-	77	0	0	0	100
SSON53G_RS20465	hypothetical protein	*yggM*	1	2	-	-	127	0	100	90.91	100
SSON53G_RS20485	membrane protein	-	2	4	fig|216599.13.peg.3328	-	130	0.0154	100	100	100
SSON53G_RS20575	acetyltransferase	*yhhY*	1	2	fig|216599.13.peg.3343	-	125	0.0160	100	100	98.70
SSON53G_RS21560	alpha-glucosidase	-	1	1	fig|216599.13.peg.3488	Carbohydrates	109	0	0	100	98.70
SSON53G_RS23905	vitamin B12 transporter BtuB	*btuB*	16	23	fig|216599.13.peg.3883	Cofactors, Vitamins, Prosthetic Groups, Pigments	23	0	31.25	21.21	14.29
SSON53G_RS24785	hypothetical protein	-	1	1	-	-	121	0	93.75	93.94	97.40
SSON53G_RS26020	LuxR family transcriptional regulator	*yjjQ*	2	2	fig|216599.13.peg.4221	-	78	0	6.25	0	100